이 학 박 사 학 위 논 문

# High-dimensional Covariance/Precision Matrix Estimation under General Missing Dependency

# 일반적인 결측 구조 하의 고차원 (역)공분산 행렬 추론

2020년 2월

서울대학교 대학원

통계학과

박 성 오

**High-dimensional Covariance/Precision Matrix Estimation under General Missing Dependency**

**일반적인 결측 구조 하의 고차원 (역)공분산 행렬 추론**

**지도교수 임 요 한**

**이 논문을 이학박사 학위논문으로 제출함**

**2020년 1월**

**서울대학교 대학원**

**통계학과**

**박 성 오**

**박성오의 이학박사 학위논문을 인준함**

**2020년 1월**

| | | |
|---|---|---|
| **위 원 장** | **이 재 용** | **(인)** |
| **부위원장** | **임 요 한** | **(인)** |
| **위 원** | **원 중 호** | **(인)** |
| **위 원** | **정 성 규** | **(인)** |
| **위 원** | **최 영 근** | **(인)** |

# High-dimensional Covariance/Precision Matrix Estimation under General Missing Dependency

By

Seongoh Park

A Thesis
Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
February, 2020

**ABSTRACT**

# High-dimensional Covariance/Precision Matrix Estimation under General Missing Dependency

Seongoh Park
The Department of Statistics
The Graduate School
Seoul National University

A sample covariance matrix $\boldsymbol{S}$ of completely observed data is the key statistic to initiate a large variety of multivariate statistical procedures, such as structured covariance/precision matrix estimation, principal component analysis, and graphical models. However, the sample covariance matrix obtained from partially observed data is not adequate to use due to its biasedness. To correct the bias, a simple adjustment method called an inverse probability weighting (IPW) has been used in previous research, yielding the IPW estimator. The estimator plays a role of $\boldsymbol{S}$ under missing data context so that it can be plugged-in into off-the-shelf multivariate procedures instead of $\boldsymbol{S}$. However, theoretical properties (e.g. concentration) of the IPW estimator have been only established under very simple structure

of missing pattern; every variable of each sample is independently subject to missing with equal probability.

We investigate the deviation of the IPW estimator when observations are partially observed under general missing dependency. We prove the optimal convergence rate $O_p(\sqrt{\log p/n})$ of the IPW estimator based on the element-wise maximum norm. We also derive similar deviation results even when implicit assumptions (known mean and/or missing probability) are relaxed. The optimal rate is especially crucial in estimating a precision matrix, because of the "meta-theorem" (Liu et al. 2012) that claims the rate of the IPW estimator governs that of the resulting precision matrix estimator. In the simulation study, we discuss non-positive semi-definiteness of the IPW estimator and compare the estimator with imputation methods, which are practically important.

**Keywords:** Convergence rate, covariance matrix, dependent missing structure, inverse probability weighting, missing data.

**Student Number:** 2013 - 22899

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

One of the overarching themes in statistic and machine learning societies is to discover complex relationships among high-dimensional variables. Out of many, the covariance matrix and its inverse matrix (the precision matrix) are arguably important statistical tools in this line of research. Hence, methodological and theoretical analyses of these statistics, such as scalability, consistency, and convergence rate, have been established by many researchers (see Section Introduction from Fan et al. (2016) for a comprehensive literature review, and references therein), because of their utility in a broad range of disciplines such as biology, geophysics, economics, public health, and social sciences. Despite much advance over decades in the estimation of a covariance/precision matrix under the high-dimensional setting, most approaches to date have been oblivious to handling missing observations. However, widespread applications have emerged in modern science where the primary interest is placed on estimating the correlation structure involved in observations subject to missing, for example, climate data (Schneider 2001), genomic studies (Cui et al. 2017; Liang et al. 2018), and

remote sensing data (Glanz and Carvalho 2018), to name a few. Even so, there has been relatively less development in both methodology and theory that deal with the (inverse) covariance estimation problem in the presence of missing data.

## 1.1 Past works on (inverse) covariance matrix estimation with missing values

Previous researches in the field of estimation of a (inverse) covariance matrix with incomplete data, though not many to our best knowledge, can be classified into two branches; the likelihood-based method and the plug-in method.

The first line of the works is the likelihood-based inference, mostly achieving the maximum likelihood estimator by the EM algorithm (or its variants) (Allen and Tibshirani 2010; Huang et al. 2007; Liang et al. 2018; Städler and Bühlmann 2012; Thai et al. 2014). In spite of individual success in covariance/precision matrix estimation when missing observations are present, the major drawback of this approach is separate development of estimating algorithms and supporting theories. That is, one considering a new proposal under this framework should put huge efforts to implement the new method for practice purposes and prove theoretical properties (e.g. consistency). Furthermore, the Gaussian assumption on observations commonly used in the likelihood inference could be restrictive in the high-dimensional setting.

The other scheme of research studied rather in recent years utilizes the idea of a plug-in estimator, based on the fact that many procedures to es-

timate a covariance/precision matrix solely rely on the sample covariance matrix $\boldsymbol{S}$, not the data itself. Preceding works (Cai and Zhang 2016; Kolar and Xing 2012; Lounici 2014; Pavez and Ortega 2019; Rao et al. 2017; Wang et al. 2014) have considered adjusting the missing proportion, or the bias that appears in the sample covariance matrix $\boldsymbol{S}_Y$ computed by partial observations (see the definition in (2.1)). The modified estimator is often referred to as an inverse probability weighting (IPW) estimator and put into a module (procedure) of the (inverse) covariance estimation. For example, Kolar and Xing (2012) plug-in the IPW estimator into the graphical lasso procedure (Friedman et al. 2008) to estimate a sparse precision matrix, while Cai and Zhang (2016) use banding, tapering, or thresholding techniques to recover a structured covariance matrix under missing data context. Wang et al. (2014) apply the CLIME method (Cai et al. 2011) to the bias-corrected rank-based correlation matrix to estimate a sparse precision matrix of a non-paranormal distribution. In the low-rank approximation problem, the IPW estimator is plugged into the matrix lasso (Rohde and Tsybakov 2011) by Lounici (2014), which is extended by Rao et al. (2017) to vector autoregressive processes. All of these works are based on one common assumption about missing; for each sample, each variable is independently subject to missing with equal (uniform) probability. Their theoretical analyses, though recovering the aimed rate $\sqrt{\log p/n}$ ($n$: the sample size, $p$: dimension), are established based on such restrictive independent assumption. In contrast, dependent (and non-uniform) missing structure has not been paid attention to until very recent year, but was initiated by Park and Lim (2019) and investigated further by Pavez and Ortega (2019). While the two results are based on the spectral norm using

the effective rank of a matrix (see Table 2.1), we derive the optimal convergence rate of the IPW estimator in terms of the element-wise maximum norm under general missing dependency.

## 1.2 Our contributions

Our main contributions are outlined here.

*Derivation of the optimal convergence rate under dependent missing structure.* We develop a non-asymptotic deviation inequality of the IPW estimator in the element-wise maximum norm by extending missing dependency (Theorem 1). The theoretical results maintain the conventional convergence rate $\sqrt{\log p/n}$ achieved by the earlier works (Bickel and Levina (2008a) and the references in Table 2.1). Theorem 1 can be further used to estimate the structured precision matrix for the Gaussian graphical model, due to no assumptions on the covariance/precision matrix, the sample size, or the dimension.

*Relaxation of implicit assumptions to derive the rates.* In analyzing the concentration of the IPW estimator, estimation of the population mean and missing probability has been largely unexplored (Lounici (2014), Wang et al. (2014), Park and Lim (2019), Pavez and Ortega (2019)), which is not desirable in practice. Filling the gaps, this thesis establishes the concentration inequalities for the IPW estimator under unknown mean (Theorem 2) and missing probability (Theorem 3).

## 1.3 Outline

The remainder of this thesis is organized as follows. At the beginning of Chapter 2, we formally state the problem setup and introduce the IPW estimator under general missing dependency. Based on it, we present our theoretical results in Chapter 2 and their variants considering relaxations in Chapter 3. Chapter 4 deals with non-positive semi-definiteness of the IPW estimator and its potential remedies. In Chapter 5 and 6, we show our numerical studies on simulated data and real data. We conclude this thesis with a brief discussion and summary in Chapter 7.

# Chapter 2

# The IPW estimator under general missing structure and its rate

Let $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a $p$-dimensional random variable with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbb{E}(XX^{\mathrm{T}})$. We denote missing observations by 0, which has a simple mathematical representation using a missing indicator[1] $\delta_j$ that takes its value either 0 (missing) or 1 (observed);

$$Y = (Y_1, \ldots, Y_p)^{\mathrm{T}}, \quad Y_j = \delta_j X_j, \quad j = 1, \ldots, p.$$

The multivariate binary vector $\delta = (\delta_1, \ldots, \delta_p)^{\mathrm{T}}$ is assumed to follow some distribution where a marginal distribution of $\delta_j$ is the Bernoulli distribution with success probability $0 \leq \pi_j \leq 1$. This formulation is an extension

[1] Technically, $\delta_j$ is a "response" indicator as termed in Kim and Shao (2013), since the value 1 indicates an observed (responded) variable, but we insist on using "missing" to emphasize the context of missing data.

of the independent missing structure used in previous works (Cai et al. 2016; Kolar and Xing 2012; Lounici 2014; Wang et al. 2014), which assume missing indicator $\delta_k$ is independent of $\delta_\ell$ $(k \neq \ell)$. Contrary to it, this thesis assumes the $p$ random variables $\{\delta_j, j = 1, \ldots, p\}$ are allowed to be dependent and not identically distributed. The probability of observing at multiple positions is henceforth denoted by

$$\mathrm{P}(\delta_i = \delta_j = \delta_k = \ldots = 1) = \pi_{ijk\ldots}.$$

Dependent structure in missing naturally occurs through a longitudinal clinical study since a patient absent at visit(=variable) $k$ would have more possibility of not showing up at forthcoming visits $\ell(> k)$. There exists more general and plausible scenario where extrinsic covariates are involved in occurrence of missing.

Let us consider $n$ samples from the population above where the covariance matrix $\mathbf{\Sigma} = (\sigma_{k\ell}, 1 \leq k, \ell \leq p)$ is to be estimated. Denote the $i$-th sample version of $X, Y, \delta_j$ by $X_i, Y_i, \delta_{ij}$, respectively. Then, the sample covariance matrix from partially observed data is obtained by

$$\boldsymbol{S}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^{\mathrm{T}} = \Big(\frac{1}{n}\sum_{i=1}^{n} \delta_{ij}\delta_{ik} X_{ij} X_{ik}, 1 \leq j, k \leq p\Big). \tag{2.1}$$

It can be easily checked that $\boldsymbol{S}_Y$ is biased for $\mathbf{\Sigma}$, since its expectation is $\mathbf{\Sigma}^{\pi} = \Big(\pi_{jk}\sigma_{jk}, 1 \leq j, k \leq p\Big)$ by assuming independence between $\{X_i\}_{i=1}^n$ and $\{\delta_{ij}\}_{i,j}$. This motivates one to adjust an weight of each component of $\boldsymbol{S}_Y$ and to define the IPW estimator $\hat{\mathbf{\Sigma}}^{IPW} = \Big((\hat{\mathbf{\Sigma}}^{IPW})_{jk}, 1 \leq j, k \leq p\Big)$ by

$$(\hat{\mathbf{\Sigma}}^{IPW})_{jk} = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \frac{\delta_{ij}}{\pi_j} X_{ij}^2 & j = k, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \frac{\delta_{ij}\delta_{ik}}{\pi_{jk}} X_{ij} X_{ik} & j \neq k, \end{cases} \tag{2.2}$$

provided that $\pi_{jk} > 0, \forall j, k$[2]. Then, $\hat{\mathbf{\Sigma}}^{IPW}$ is unbiased for $\mathbf{\Sigma}$ under the missing completely at random (MCAR) mechanism (Little and Rubin 1986), that is, $\{\delta_{ij}\}_{j=1}^{p}$ is independent of $\{X_{ij}\}_{j=1}^{p}$ for $i = 1, \ldots, n$. For example, when data acquisition is carried out through censors (e.g. remote sensing data), missing arises due to faults in censors and thus is independent of values to be measured.

We note this adjustment technique is frequently used in general context of missing data and also known as the propensity score method. The underlying idea of it is to construct an unbiased estimating equation by reweighting the contribution of each sample on the equation. The corresponding equation for the covariance estimation problem under the Gaussian setting without missing is a score function given by

$$\frac{1}{n}\sum_{i=1}^{n} Q(X_i; \mathbf{\Sigma}) = 0, \tag{2.3}$$

where $Q(X_i; \mathbf{\Sigma}) = \mathbf{\Sigma}^{-1} X_i X_i^{\mathrm{T}} \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}$. Since (2.3) is equivalent to solve $n^{-1}\sum_{i=1}^{n}(X_i X_i^{\mathrm{T}} - \mathbf{\Sigma}) = 0$, the reweighted version of the equation above would be

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{R}_i * (X_i X_i^{\mathrm{T}} - \mathbf{\Sigma}) = 0, \tag{2.4}$$

where $\boldsymbol{R}_i = \left(\delta_{ij}\delta_{ik}/\pi_{jk}, 1 \leq j, k \leq p\right)$ and $*$ is an element-wise product. Solving the equation above with respect to $\mathbf{\Sigma}$ yields an empirical version of the IPW estimator that replaces $\pi_{jk}$ in (2.2) with $n^{-1}\sum_{i=1}^{n}\delta_{ij}\delta_{ik}$. This estimator has been used and analyzed before in Kolar and Xing (2012) and Cai and Zhang (2016), which will be studied in Chapter 3.2 of this thesis under general missing dependency. Remark that the inverse probability $\pi_{jk}$ in $\boldsymbol{R}_i$ at (2.4) is ignorable and does not play any role in defining the

[2]By definition, $\pi_{jj} = \pi_j$

empirical estimator. However, when the probability is dependent on sample-specific variables ($X_i$ or extrinsic covariates $W_i$), we should give weights in the form of the conditional probability defined by $\mathrm{P}(\delta_{ij} = \delta_{ik} = 1 | X_i, W_i)$, which adjusts the selection bias from partial observations $\{i : \delta_{ij} = \delta_{ik} = 1\}$. For the sake of simplicity, analyses in this thesis only concern the identical setting on missing indicators, that is, $\pi_{jk\ell\ldots} \overset{\forall i}{=} \mathrm{P}(\delta_{ij} = \delta_{ik} = \delta_{i\ell} = \ldots = 1)$.

## 2.1 Notations

Through out this thesis, we will use the following matrix norms; for a matrix $\boldsymbol{A}$, the element-wise maximum norm is $||\boldsymbol{A}||_{max} = \max_{i,j} |\boldsymbol{A}_{ij}|$, the operator 1-norm is $||\boldsymbol{A}||_1 = \max_j \sum_i |\boldsymbol{A}_{ij}|$, the operator 2-norm $||\boldsymbol{A}||_2$ is the largest singular value (or eigenvalue if $\boldsymbol{A}$ is symmetric), the element-wise 1-norm is $|\boldsymbol{A}|_1 = \sum_{i,j} |\boldsymbol{A}_{ij}|$, and the Frobenius norm is $||\boldsymbol{A}||_F = \sqrt{\sum_{i,j} \boldsymbol{A}_{ij}^2}$. $\mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix whose diagonal entries are inherited from $\boldsymbol{A}$. For a vector $v$, we define $||v||_1 = \sum_j |v_j|$. Also, we define $R(\theta) = \exp(1/(4e\theta^2)) - 1/2 - 1/(4e\theta^2)$ $(\theta > 0)$ which is monotonically decreasing and satisfies $R(\theta) > 1/2$.

## 2.2 Assumptions

We state our assumptions used in the following theoretical analyses; (i) sub-Gaussianity for each component of $X_i$, (ii) a general dependency structure for $\delta_i$, and (iii) MCAR for missing mechanism. We begin with one of the equivalent definitions of the sub-Gaussian variable (Vershynin 2018); the uniformly bounded moments.

**Assumption 1** (Sub-Gaussianity)**.** *$X$ is a sub-Gaussian random variable in $\mathbb{R}$ satisfying*

$$\mathbb{E}X = 0, \quad \mathbb{E}X^2 = 1, \qquad \textit{and } \sup_{r \geq 1} \frac{\left\{\mathbb{E}|X|^r\right\}^{1/r}}{\sqrt{r}} \leq K \tag{2.5}$$

*for some $K > 0$.*

We note that the Gaussian random variable $X \sim N(0, \sigma^2)$ satisfies

$$\sup_{r \geq 1} \left\{\mathbb{E}|X|^r\right\}^{1/r} / \sqrt{r} \leq \sigma K$$

for some numeric constant $K > 0$.

Missing is assumed to occur with dependency in sense of the following;

**Assumption 2** (General missing dependency)**.** *A missing indicator $\delta = (\delta_1, \ldots, \delta_p)^{\mathrm{T}} \in \{0, 1\}^p$ follows some multivariate distribution where each marginal distribution is a Bernoulli distribution with a missing probability*[3] *$\pi_j \in (0, 1]$, i.e., $\delta_j \sim Ber(\pi_j)$. Further assume that $\pi_{jk} \neq 0$ for all $1 \leq j, k \leq p$.*

The non-degenerate condition for the missing probabilities (i.e., $\pi_j > 0, \pi_{jk} > 0$) is required since, for example, $\pi_{jk} = 0$ implies no data could be observed for estimating the second moment $\sigma_{jk}$, which is unrealistic for our discussion. Next, we formally state our missing mechanism again;

**Assumption 3** (Missing completely at random)**.** *An event that an observation is missing is independent with both observed and unobserved random variables.*

Under the data structure in this thesis, the above mechanism essentially says that two random vectors, $\delta_i$ and $X_i$, are independent. We note that

[3] Following the previous footnote, $\pi_{k\ldots}$ is called a "missing" probability.

Assumption 1 and 3 are commonly used in the context of covariance estimation with incomplete data, while Assumption 2 is more general than the independent structure previous researches depend on.

## 2.3 Preliminary results

We first introduce concentration inequalities of independent sum of (squared) sub-Gaussian variables that our theoretical analyses are based on.

**Lemma 1.** *Assume that $X$ is a random variables satisfying Assumption 1 for some $K > 0$. Then, the i.i.d copies $X_1, \ldots, X_n$ of $X$ satisfy,*

*(a) for $0 \le x \le eK + (2eK)^{-1}$,*

$$\mathrm{P}\left[\left|\sum_{j=1}^{n} X_j\right| \ge nx\right] \le 2\exp\left\{-\frac{nx^2}{8(1/2 + K^2e^2)}\right\},$$

*(b) and for $0 \le x \le 4eK^2R(K)$,*

$$\mathrm{P}\left[\left|\sum_{j=1}^{n}(X_j^2 - 1)\right| \ge nx\right] \le 2\exp\left\{-\frac{nx^2}{16(4eK^2)^2R(K)}\right\},$$

*where $R(t) = \exp\{1/(4et^2)\} - 1/2 - 1/(4et^2)$, $t > 0$.*

*Proof.* The proofs of (a) and (b) directly come from applications of Lemma 2 and 3. □

The first supporting lemma tells a tail bound of a variable with a cumulant generating function dominated by a quadratic function.

**Lemma 2** (Theorem 3.2 and Lemma 2.4 in Saulis and Statulevičius (1991))**.** *Let a random variable $\xi_j$ with $\mathbb{E}\xi_j = 0, \mathrm{Var}(\xi_j) = \sigma_j^2$ satisfy the following; there exist positive constants $A, C, c_1, c_2, \ldots,$ such that*

$$\left|\log \mathbb{E}\exp\{\lambda\xi_j\}\right| \le c_j^2\lambda^2, \quad |\lambda| < A, \quad \forall j, \tag{2.6}$$

*and*

$$\overline{\lim_{n\to\infty}} \sum_{j=1}^{n} c_j^2 \Big/ \sum_{j=1}^{n} \sigma_j^2 \le C.$$

*Then, we have for* $\xi = \sum_{j=1}^{n} \xi_j \Big/ \sqrt{\sum_{j=1}^{n} \sigma_j^2}$,

$$\mathrm{P}\big[\pm\xi \ge x\big] \le \exp\big(-x^2/8C\big), \quad 0 \le x \le 2AC\sqrt{\sum_{j=1}^{n} \sigma_j^2}.$$

*Furthermore, if* $\xi_i$*'s are identically distributed and satisfying the conditions above, then the variance term* $\sigma_j^2$ *does not appear in the concentration inequality:*

$$\mathrm{P}\bigg[\pm\sum_{j=1}^{n} \xi_j \ge x\bigg] \le \exp\bigg\{-\frac{x^2}{8nc_1^2}\bigg\}, \quad 0 \le x \le 2Anc_1^2.$$

The following auxiliary results for a sub-Gaussian variable $X$ facilitate one to check the condition (2.6) in Lemma 1.

**Lemma 3.** *Assume that* $X$ *is a random variables satisfying Assumption 1 for some* $K > 0$*. Then, it holds*

*(a) for* $|t| \le (2eK)^{-1}$,

$$\mathbb{E}\exp(tX) \le \exp\big\{(1/2 + K^2e^2)t^2\big\},$$

*(b) and for* $|t| < 1/(2\kappa)$,

$$\mathbb{E}\Big[\exp\big\{t(X^2 - 1)\big\}\Big] \le \exp(c_0t^2),$$

*where* $\kappa = 4eK^2$ *and* $c_0 = 2\kappa^2\{\exp(1/\kappa) - 1/2 - 1/\kappa\}$.

*Proof.* We first prove (a). For $t \in \mathbb{R}$, observe that

$$\begin{aligned}
\mathbb{E}\exp(tX) &= 1 + \frac{t^2}{2} + \sum_{r\geq 3} \frac{\mathbb{E}X^r t^r}{r!} \\
&\leq 1 + \frac{t^2}{2} + \sum_{r\geq 3} \frac{\mathbb{E}|X|^r t^r}{r!} \\
&\leq 1 + \frac{t^2}{2} + \sum_{r\geq 3} \frac{K^r r^r |t|^r}{r!} \\
&\leq 1 + \frac{t^2}{2} + \sum_{r\geq 3} K^r e^r |t|^r \ (\because (r/e)^r \leq r!) \\
&\leq 1 + \frac{t^2}{2} + \frac{(Ke|t|)^3}{1 - Ke|t|}, \quad \text{if } |t| \leq (Ke)^{-1}.
\end{aligned}$$

Then, it holds for any $0 < t_0 < (Ke)^{-1}$ that for all $|t| < t_0$,

$$\mathbb{E}\exp(tX) \leq 1 + |t|^2(1/2 + K^2e^2) \leq \exp\left\{|t|^2(1/2 + K^2e^2)\right\},$$

which concludes the proof of (a).

Next, we prove (b). Using the Minkowski inequality, we have

$$\left(\mathbb{E}|X^2 - 1|^r\right)^{1/r} \leq \left(\mathbb{E}|X|^{2r}\right)^{1/r} + 1 \leq 2rK^2 + 1,$$

which thus gives the upper bound of moments of $X^2 - 1$,

$$\mathbb{E}|X^2 - 1|^r \leq (2rK^2 + 1)^r \leq 2^{r-1}(2^r r^r K^{2r} + 1).$$

Therefore,

$$\begin{aligned}
\mathbb{E}\Big[\exp\big\{t(X^2-1)\big\}\Big] &= 1+\sum_{r\geq 2}\frac{t^r\mathbb{E}(X^2-1)^r}{r!} \\
&\leq 1+\sum_{r\geq 2}\frac{|t|^r 2^{r-1}(2^r r^r K^{2r}+1)}{r!} \\
&\leq 1+\frac{1}{2}\sum_{r\geq 2}\Big\{\frac{(4|t|rK^2)^r}{r!}+\frac{(2|t|)^r}{r!}\Big\} \\
&\leq 1+\frac{1}{2}\sum_{r\geq 2}\Big\{(4|t|eK^2)^r+\frac{(2|t|)^r}{r!}\Big\} \\
&= 1+\frac{|t|^2}{2}\sum_{r\geq 2}\Big\{(4eK^2)^2(4|t|eK^2)^{r-2}+\frac{4(2|t|)^{r-2}}{r!}\Big\}
\end{aligned}$$

where the last inequality is derived from $(n/e)^n \leq n!$ for $n \geq 1$. Then, it holds for any $0 < t_0 < 1/(4eK^2)$ that for all $|t| < t_0$,

$$\mathbb{E}\Big[\exp\big\{t(X^2-1)\big\}\Big] \leq 1+ct^2 \leq \exp(ct^2)$$

where $c$ is a function of $t_0$ defined by

$$c = c(t_0) = \frac{1}{2}\sum_{r\geq 0}\Big\{(4eK^2)^2(4t_0eK^2)^r+\frac{4(2t_0)^r}{(r+2)!}\Big\}.$$

Calculus of infinite series at the choice of $t_0 = 1/(8eK^2)$ gives

$$c(t_0) = \frac{\exp(2t_0)-1/2-2t_0}{2t_0^2},$$

which concludes the proof of (b). □

## 2.4 Main results

Lemma 4 describes the element-wise deviation of the IPW estimator from a true covariance matrix.

**Lemma 4.** *Let $X_i \in \mathbb{R}^p$ be an i.i.d. random vector with mean 0 and covariance $\mathbf{\Sigma}$, $i = 1, \ldots, n$. Suppose the scaled random variable $X_{ik}/\sqrt{\sigma_{kk}}$ satisfies Assumption 1 with a constant $K > 0$ for all $k$. Also, let $\delta_i$ be an i.i.d. binary random vector satisfying Assumption 2, $i = 1, \ldots, n$. By observing samples $\{Y_i\}_{i=1}^n$ under Assumption 3, we have*

$$\mathrm{P}\left[n^{-1}\Big|\sum_{i=1}^{n}\Big(\frac{Y_{ik}Y_{i\ell}}{\pi_{k\ell}} - \sigma_{k\ell}\Big)\Big| \geq \frac{C(\sigma_{kk}\sigma_{\ell\ell})^{1/2}K^2R(K)^{1/2}}{\pi_{k\ell}}t\right] \leq 4\exp(-nt^2), \tag{2.7}$$

*if $t \geq 0$ satisfies*

$$\begin{cases} t^2 \leq cR\Big(\dfrac{2K}{\sqrt{\pi_k + \pi_\ell - 2\pi_{k\ell}|\rho_{k\ell}|}}\Big), & \text{if } k \neq \ell, \\ t^2 \leq cR\big(K/\sqrt{\pi_k}\big), & \text{if } k = \ell, \end{cases}$$

*where $c, C > 0$ are numerical constants and $\rho_{k\ell} = \sigma_{k\ell}/\sqrt{\sigma_{kk}\sigma_{\ell\ell}}$.*

*Proof.* Assume $k$ and $\ell$ are distinct. We start by decoupling the product of two sub-Gaussian variables $Y_{ik}Y_{i\ell}/\pi_{k\ell}$ using an identity $xy = \{(x+y)^2 - (x-y)^2\}/4$ so that we have for $t \geq 0$,

$$\begin{aligned} &\left\{\Big|\sum_{i=1}^{n}\Big(\frac{Y_{ik}Y_{i\ell}}{\pi_{k\ell}} - \sigma_{k\ell}\Big)\Big| \geq nt\right\} \\ \subset\ &\left\{\Big|\sum_{i=1}^{n}\Big\{(Y_{ik}^* + Y_{i\ell}^*)^2 - \mathbb{E}(Y_{ik}^* + Y_{i\ell}^*)^2\Big\}\Big| \geq \frac{2n\pi_{k\ell}t}{\sqrt{\sigma_{kk}\sigma_{\ell\ell}}}\right\} \\ &\cup\left\{\Big|\sum_{i=1}^{n}\Big\{(Y_{ik}^* - Y_{i\ell}^*)^2 - \mathbb{E}(Y_{ik}^* - Y_{i\ell}^*)^2\Big\}\Big| \geq \frac{2n\pi_{k\ell}t}{\sqrt{\sigma_{kk}\sigma_{\ell\ell}}}\right\} \end{aligned} \tag{2.8}$$

where $Y_{ik}^* = Y_{ik}/\sqrt{\sigma_{kk}}$. Let $v_{k\ell} = \mathbb{E}|Y_{ik}^* + Y_{i\ell}^*|^2 = \pi_k + \pi_\ell + 2\pi_{k\ell}\rho_{k\ell}$. To apply Lemma 1 in Appendix, we first show $Y_{ik}^* + Y_{i\ell}^*$ is a sub-Gaussian variable satisfying the conditions of the lemma.

**Lemma.** *For $i = 1, \ldots, n$ and $1 \le k \ne \ell \le p$, we have*

$$\sup_{r \ge 1} \frac{\{\mathbb{E}|Y_{ik} + Y_{i\ell}|^r\}^{1/r}}{\sqrt{r v_{k\ell}}} \le 2K/\sqrt{v_{k\ell}}.$$

*Proof.* To obtain an uniform bound on higher moments, we observe that

$$\begin{aligned}
\frac{\{\mathbb{E}|Y^*_{ik} + Y^*_{i\ell}|^r\}^{1/r}}{\sqrt{r}} &\le \frac{2^{1-1/r}\{\mathbb{E}|Y^*_{ik}|^r + \mathbb{E}|Y^*_{i\ell}|^r\}^{1/r}}{\sqrt{r}} \\
&= \frac{2^{1-1/r}\Big\{\pi_k \mathbb{E}\big|X_{ik}/\sqrt{\sigma_{kk}}\big|^r + \pi_\ell \mathbb{E}\big|X_{i\ell}/\sqrt{\sigma_{\ell\ell}}\big|^r\Big\}^{1/r}}{\sqrt{r}} \\
&\le \frac{2^{1-1/r}\Big\{\pi_k(\sqrt{r}K)^r + \pi_\ell(\sqrt{r}K)^r\Big\}^{1/r}}{\sqrt{r}} \\
&\le 2K\Big(\frac{\pi_k + \pi_\ell}{2}\Big)^{1/r}
\end{aligned}$$

where the first inequality holds due to convexity of $x \mapsto |x|^r (r \ge 1)$ and the third inequality uses the moment condition of the sub-Gaussian variable $X_{ik}/\sqrt{\sigma_{kk}}$. We note that $\Big(\frac{\pi_k + \pi_\ell}{2}\Big)^{1/r} \le 1$ for all $r \ge 1$ since $0 \le (\pi_k + \pi_\ell)/2 \le 1$. which concludes the proof. □

By applying Lemma 1 (b), we have for some numerical constants $c, C > 0$,

$$\mathrm{P}\Big[\Big|\sum_{i=1}^n \Big\{(Y^*_{ik}+Y^*_{i\ell})^2 - v_{k\ell}\Big\}\Big| \ge \frac{2n\pi_{k\ell}t}{\sqrt{\sigma_{kk}\sigma_{\ell\ell}}}\Big] \le 2\exp\Big\{-\frac{Cn\pi^2_{k\ell}t^2}{K^4\sigma_{kk}\sigma_{\ell\ell}R(2K/\sqrt{v_{k\ell}})}\Big\},$$

for $0 \le t \le \dfrac{c(\sigma_{kk}\sigma_{\ell\ell})^{1/2}K^2R(2K/\sqrt{v_{k\ell}})}{\pi_{k\ell}}$. Hence, replacing $t$ by

$$\tilde{t} \equiv \frac{(\sigma_{kk}\sigma_{\ell\ell})^{1/2}K^2R(2K/\sqrt{v_{k\ell}})^{1/2}}{C^{1/2}\pi_{k\ell}}t, \quad t > 0,$$

in the above inequality, we get

$$\mathrm{P}\Big[\Big|\sum_{i=1}^n \Big\{(Y_{ik} + Y_{i\ell})^2 - \mathbb{E}(Y_{ik} + Y_{i\ell})^2\Big\}\Big| \ge 2n\pi_{k\ell}\tilde{t}\Big] \le 2\exp\{-nt^2\},$$

if $0 \le t \le \tilde{c}\sqrt{R(2K/\sqrt{v_{k\ell}})}$ for some numerical constant $\tilde{c} > 0$. Note that

$$R\left(\frac{2K}{\sqrt{\pi_k + \pi_\ell - 2\pi_{k\ell}|\rho_{k\ell}|}}\right) \le R\left(\frac{2K}{\sqrt{v_{k\ell}}}\right) \le R(K),$$

and using this bounds, we now have

$$\mathrm{P}\left[\Big|\sum_{i=1}^{n}\Big\{(Y_{ik}+Y_{i\ell})^2 - \mathbb{E}(Y_{ik}+Y_{i\ell})^2\Big\}\Big| \ge 2n\pi_{k\ell}\frac{(\sigma_{kk}\sigma_{\ell\ell})^{1/2}K^2R(K)^{1/2}}{C^{1/2}\pi_{k\ell}}t\right] \le 2\exp\{-nt^2\},$$

for $0 \le t \le \tilde{c}\sqrt{R\left(\frac{2K}{\sqrt{\pi_k + \pi_\ell - 2\pi_{k\ell}|\rho_{k\ell}|}}\right)}$. The similar statement holds with $Y_{ik} - Y^*_{i\ell}$. Therefore, combining these results with (2.8) yield

$$\mathrm{P}\left[n^{-1}\Big|\sum_{i=1}^{n}\Big(\frac{Y_{ik}Y_{i\ell}}{\pi_{k\ell}} - \sigma_{k\ell}\Big)\Big| \ge \frac{(\sigma_{kk}\sigma_{\ell\ell})^{1/2}K^2R(K)^{1/2}}{C^{1/2}\pi_{k\ell}}t\right] \le 4\exp\{-nt^2\},$$

for $0 \le t \le \tilde{c}\sqrt{R\left(\frac{2K}{\sqrt{\pi_k + \pi_\ell - 2\pi_{k\ell}|\rho_{k\ell}|}}\right)}$, which completes the proof for the case of $k \neq \ell$.

The concentration inequality for diagonal entries (i.e., $k = \ell$) of the IPW estimate is similarly derived. One can easily check

$$\sup_{r\ge 1}\frac{\{\mathbb{E}|Y_{ik}|^r\}^{1/r}}{\sqrt{r\pi_k\sigma_{kk}}} \le K/\sqrt{\pi_k}.$$

Then, due to Lemma 1 (b), we get

$$\mathrm{P}\left[n^{-1}\Big|\sum_{i=1}^{n}\Big(\frac{Y_{ik}^2}{\pi_k} - \sigma_{kk}\Big)\Big| \ge \frac{\tilde{C}\sigma_{kk}K^2R(K)^{1/2}}{\pi_k}t\right] \le 2\exp\{-nt^2\},$$

for $0 \le t \le \sqrt{R(K/\sqrt{\pi_k})}$. This concludes the whole proof. □

We provide some remarks regarding to this lemma. This concentration inequality covers the existing results as special cases. First, if data is assumed to be fully observed (i.e., $\pi_{k\ell} = 1, \forall k, \ell$), then (2.7) is reduced to

$$\mathrm{P}\left[n^{-1}\Big|\sum_{i=1}^{n}\big(X_{ik}X_{i\ell} - \sigma_{k\ell}\big)\Big| \ge C_1\sqrt{\sigma_{kk}\sigma_{\ell\ell}}\,t\right] \le C_2\exp(-nt^2), \quad 0 \le t \le C_3,$$

where $C_1, C_2, C_3$ are numerical constants. It can be seen that this form is equivalent to Lemma A.3. in Bickel and Levina (2008b) (Gaussian) or Lemma 1 in Ravikumar et al. (2011) (sub-Gaussian), up to multiple constant difference. When independent and identical structure of missing indicators is assumed (i.e., $\delta_k \overset{\forall k}{\sim} \mathrm{Ber}(\pi)$) in Lemma 4, the reduced probabilistic bound is similar to that from Kolar and Xing (2012) (plugging in $t \leftarrow \sqrt{\log(4/\delta)/n}$ in (2.7))

$$\mathrm{P}\left[n^{-1}\Big|\sum_{i=1}^{n}\Big(\frac{Y_{ik}Y_{i\ell}}{\pi^2}-\sigma_{k\ell}\Big)\Big| \geq \frac{CK^2}{\pi^2}\sqrt{\frac{R(K)\sigma_{kk}\sigma_{\ell\ell}\log(4/\delta)}{n}}\right] \leq \delta$$

for the sample size $n$ chosen according to Lemma 4. Rigorously speaking, the proposed IPW estimator in Lemma 4 and that of Kolar and Xing (2012) (see (3.5)) are different by the inverse weighting factor when correcting missing observations. However, replacing missing probabilities with unbiased empirical estimates will not cause a considerable change in our result (see Chapter 3.2).

Using the lemma above, the rate of convergence of the IPW estimate can be derived in terms of the element-wise maximum norm. Let us define the maximum and minimum value of parameters that appear in Lemma 4 as follows;

$$\sigma_{max} = \max_{k} \sigma_{kk}, \quad \pi_{min} = \min_{k,\ell} \pi_{k\ell},$$

$$v_{min} = \min_{k\neq\ell}(\pi_k + \pi_\ell - 2\pi_{k\ell}|\rho_{k\ell}|).$$

**Theorem 1.** *Assume the conditions of Lemma 4, and further assume the sample size and dimension satisfy*

$$n/\log p > c\left\{\exp\left(\frac{v_{min}}{16eK^2}\right) - \frac{1}{2} - \frac{v_{min}}{16eK^2}\right\}^{-1} \tag{2.9}$$

*then it holds that*

$$\mathrm{P}\left[\left|\left|\hat{\mathbf{\Sigma}}^{IPW} - \mathbf{\Sigma}\right|\right|_{max} \geq \frac{C\sigma_{max}K^2}{\pi_{min}}\sqrt{\frac{R(K)\log p}{n}}\right] \leq 4p^{-1},$$

*where $c, C > 0$ are numerical constants.*

*Proof.* From Lemma 4, it holds that for $1 \leq k, \ell \leq p$,

$$\mathrm{P}\left[n^{-1}\Big|\sum_{i=1}^{n}\Big(\frac{Y_{ik}Y_{i\ell}}{\pi_{k\ell}} - \sigma_{k\ell}\Big)\Big| \geq \frac{C\sigma_{max}K^2R(K)^{1/2}}{\pi_{min}}t\right] \leq 4\exp(-nt^2),$$

if $t \geq 0$ satisfies, since $R$ is monotonically decreasing,

$$\begin{cases} t^2 \leq cR\big(2K/\sqrt{v_{min}}\big), & \text{if } k \neq \ell, \\ t^2 \leq cR\big(K/\sqrt{\pi_{min,d}}\big) & \text{if } k = \ell, \end{cases}$$

where $\pi_{min,d} = \min_k \pi_k$. Then, by using an union bound argument, we get

$$\mathrm{P}\left[\max_{k,\ell}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{Y_{ik}Y_{i\ell}}{\pi_{k\ell}} - \sigma_{k\ell}\Big)\Big| \geq \frac{C\sigma_{max}K^2\sqrt{R(K)}\,t}{\pi_{min}}\right] \quad \leq \quad 4p^2\exp(-nt^2).$$

for $t^2/c \leq R\big(K/\sqrt{(v_{min}/4) \wedge \pi_{min,d}}\big) = R\big(2K/\sqrt{v_{min}}\big)$. Note that $v_{min}/4 \leq \pi_{min,d}$.

Then, by plugging-in $t \leftarrow \alpha\sqrt{\log p/n}$ ($\alpha > 0$), we get the convergence rate of the maximum norm of the IPW estimate

$$\mathrm{P}\left[\max_{k,\ell}\big|(\hat{\mathbf{\Sigma}}^{IPW})_{k\ell} - \sigma_{k\ell}\big| \geq \frac{C\sigma_{max}K^2\alpha}{\pi_{min}}\sqrt{\frac{R(K)\log p}{n}}\right] \quad \leq \quad 4p^{2-\alpha^2},$$

if $0 \leq \alpha^2 \leq cR\big(2K/\sqrt{v_{min}}\big)n/\log p$. Suppose $n, p$ satisfy

$$n/\log p > \frac{9}{c^2R\big(2K/\sqrt{v_{min}}\big)}$$

so that we can choose $\alpha^2 = 3$. This concludes the proof. □

If we assume the independent structure on missing indicators, we get the following result, which is comparable to those from Kolar and Xing (2012) and Lounici (2014). Let $\hat{\mathbf{\Sigma}}_{ind}^{IPW}$ be the IPW estimator (2.2) with $\pi_{jk} = \pi^2, j \neq k$ and $\pi_{jj} = \pi$ for all $j, k$.

**Corollary 1** (Identical and independent missing structure)**.** *Under the conditions of Lemma 4, we further assume $\delta_{ik} \sim Ber(\pi)$, independently, $k = 1, \ldots, p$. Then, when the sample size and dimension satisfy*

$$n/\log p > c\Big\{ \exp\Big(\frac{\pi(1-\pi\rho_{max})}{8eK^2}\Big) - \frac{1}{2} - \frac{\pi(1-\pi\rho_{max})}{8eK^2} \Big\}^{-1},$$

*then it holds that*

$$\mathrm{P}\left[ \big|\big|\hat{\mathbf{\Sigma}}_{ind}^{IPW} - \mathbf{\Sigma}\big|\big|_{max} \geq \frac{C\sigma_{max}K^2}{\pi^2}\sqrt{\frac{R(K)\log p}{n}} \right] \leq 4p^{-1},$$

*where $c, C > 0$ are numerical constants and $\rho_{max} = \max_{k\neq\ell} |\rho_{k\ell}|$.*

Note that the Taylor expansion of an exponential function yields

$$\Big\{ \exp\Big(\frac{\pi(1-\pi\rho_{max})}{8eK^2}\Big) - \frac{1}{2} - \frac{\pi(1-\pi\rho_{max})}{8eK^2} \Big\}^{-1} = c_1/\big(1 + c_2\pi^2 + o(\pi^2)\big), \quad \text{as } \pi \to 0,$$

for some $c_1, c_2 > 0$. Therefore, the sample size (relative to the dimension) required for accurate estimation is less sensitive to the missing probability $\pi$ compared to the previous works (Kolar and Xing 2012; Lounici 2014; Wang et al. 2014) whose magnitude is in order of $1/\pi^2$ (see Table 2.1). However, the bound of the IPW estimator in the element-wise maximum norm increases in the order of magnitude $1/\pi^2$, which is larger than the rate $1/\pi$ claimed in other literature (see Table 2.1).

## 2.5 Comparison of the rates

Table 2.1 summarizes the rate and sample size of the IPW estimator from the related works. Cai and Zhang (2016) have considered the minimax op-

timality (with a structured covariance matrix), which is, however, not comparable to what is given in Table 2.1. Hence, their work is not included here.

| Article | Est. | Norm | Rate | Size |
|---|---|---|---|---|
| K2012 | $\hat{\mathbf{\Sigma}}^{emp}$ (3.5) | $\|\|\cdot\|\|_{max}$ | $\sigma_{max}\sqrt{\frac{\log(8p)}{\pi^2 n - \sqrt{2\pi^2 n \log(2p)}}}$ | $p = O(\exp(n\pi^2))$ |
| W2014 | Spearman's $\rho$ | $\|\|\cdot\|\|_{max}$ | $\sqrt{\frac{\log p}{\pi^2 n}}$ | $p = O(\exp(n\pi^2))$ |
| W2014 | Kendall's $\tau$ | $\|\|\cdot\|\|_{max}$ | $\sqrt{\frac{\log p}{\pi^2 n}}$ | $p = O(\exp(n\pi^2))$ |
| L2014 | $\hat{\mathbf{\Sigma}}^{IPW}_{ind}$ | $\|\|\cdot\|\|_2$ | $\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})\|\|\mathbf{\Sigma}\|\|_2 \log p}{\pi^2 n}}$ | $p = O(\exp(n\pi^2\|\|\mathbf{\Sigma}\|\|_2/\mathrm{tr}(\mathbf{\Sigma})))$ |
| PL2019 | $\hat{\mathbf{\Sigma}}^{IPW}$ (2.2) | $\|\|\cdot\|\|_2$ | $\|\|M\|\|_2\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})\|\|\mathbf{\Sigma}\|\|_2 \log p}{n}}$ | $p = O\Big(\exp(\{n\|\|\mathbf{\Sigma}\|\|_2/\mathrm{tr}(\mathbf{\Sigma})\}^{1/3})\Big)$ |
| PO2019 | $\hat{\mathbf{\Sigma}}^{IPW}$ (2.2) | $\sqrt{\mathbb{E}\|\|\cdot\|\|_2^2}$ | $\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})\|\|\mathbf{\Sigma}\|\|_2 \log p}{\pi_{min} n}}$ | $p = O\Big(\exp(n\pi_{min}\|\|\mathbf{\Sigma}\|\|_2/\{(\log n)^2\mathrm{tr}(\mathbf{\Sigma})\})\Big)$ |
| Theorem 1 | $\hat{\mathbf{\Sigma}}^{IPW}$ (2.2) | $\|\|\cdot\|\|_{max}$ | $\sigma_{max}\sqrt{\frac{\log p}{\pi^2_{min} n}}$ | See (2.9) |

Table 2.1: Summary of literature using the idea of the IPW estimator. "Rate" is the convergence rate (up to a constant depending only on distributional parameters) of an estimator ("Est.") measured by a matrix norm ("Norm"). "Size" is a condition for $n$ and $p$ to guarantee the rate holds with probability at least $1/p$. At the first column, we use the following labels: L2014=Lounici (2014), KX2012=Kolar and Xing (2012), W2014=Wang et al. (2014), PL2019=Park and Lim (2019), and PO2019=Pavez and Ortega (2019). The last three rows have considered dependency across missing indicators. $M = (1/\pi_{k\ell}, 1 \le k, \ell \le p)$.

Table 2.1 shows the rate of convergence $\sqrt{\log p/n}$ has appeared in the previous literature. When dependency for missing indicators is considered,

the achieved rate in Park and Lim (2019) under the spectral norm is not optimal, though they have first tackled it. Very recently, Pavez and Ortega (2019) show an improved rate for expectation of an estimation error based on the spectral norm. In terms of the element-wise maximum norm, to our best knowledge, this thesis is among the first to get the optimal rate.

## 2.6 The meta-theorem in estimation of a precision matrix

The derived concentration inequality is very crucial because of its application to precision matrix estimation. The related theory known as the meta-theorem that has first appeared in Liu et al. (2012) implies that the rates of the precision matrix estimator $\hat{\boldsymbol{\Omega}}$ are determined by the rate $||\cdot||_{max}$ of an input matrix (e.g. the IPW estimator) used to estimate $\hat{\boldsymbol{\Omega}}$. Therefore, when there is no missing, success of the graphical lasso (Ravikumar et al. 2011), the CLIME (Cai et al. 2011), and the graphical Dantzig selector (Yuan 2010) in accurate estimation and graph recovery depends on the fact that the sample covariance matrix $\boldsymbol{S}$ satisfies

$$\mathrm{P}\Big(\big|\big|\boldsymbol{S}-\boldsymbol{\Sigma}\big|\big|_{max} \geq C\sqrt{\log p/n}\Big) \leq d/p, \tag{2.10}$$

for some $C, d > 0$. To grasp the underlying mechanism of the meta-theorem, we refer readers to the proof of Corollary 2. Since the claimed rate of convergence in Theorem 1 is the same as that of $\boldsymbol{S}$ in (2.10), the meta-theorem also guarantees the same optimal rates of the precision matrix estimators with missing observations.

It should be remarked that the rate in Theorem 1 is not driven under some class of matrices (e.g. sparse or low-rank) for covariance/precision

matrix and under some restriction on $n$ and $p$ such as an asymptotic ratio between them, i.e., $p/n \to \alpha \in [0, \infty)$. Such flexibility makes it possible to adopt different conditions (on $\mathbf{\Sigma}$, $\mathbf{\Omega}$, $n$, or $p$) required from different precision matrix estimation methods (e.g. the graphical lasso). We describe the meta-theorem under the dependent missing structure below, which is an extension of Theorem 4.3 in Liu et al. (2012).

**Corollary 2.** *Let the true covariance matrix $\mathbf{\Sigma}$ satisfy the same assumptions that the precision matrix estimation procedure such as the graphical lasso, the graphical Dantzig selector, and the CLIME requires to guarantee the consistency and support recovery of a graph.*

*If we plug the IPW estimator $\hat{\mathbf{\Sigma}}^{IPW}$ into one of the aforementioned methods, the end product retrieves the optimal rate of convergence, and thus has consistency and support recovery properties*[4] *even under general missing dependency.*

*Proof.* We summarize theorems/lemmas from the original works that bridge the rate of the plug-in estimator with those of the final precision matrix. If $\delta = \sqrt{\log p/n}$ in each theorem, then the rates of the precision matrix are optimal and guarantee both estimation consistency in different norms and support recovery ($\because ||\cdot||_{max}$). As usual, $\hat{\mathbf{\Sigma}}^{plug}$ denotes the plug-in estimator.

### Graphical lasso

Suppose $S \subset [p] \times [p]$ is an union of a true edge set and diagonal elements. Define $\mathbf{\Gamma} = \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}$,

$$\mathbf{\Gamma}_{SS} = \left(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}\right)_{SS} = \mathbf{\Omega}_S^{-1} \otimes \mathbf{\Omega}_S^{-1},$$

[4] The support recovery is not guaranteed with the graphical Dantzig selector, since its rate is achieved in the matrix $\ell_1$-norm, not $||\cdot||_{max}$.

and similarly $\mathbf{\Gamma}_{eS} = \left(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}\right)_{eS}$, $e \in S^c$. Also, denote $\kappa_{\mathbf{\Sigma}} = ||\mathbf{\Sigma}||_\infty$ and $\kappa_{\mathbf{\Gamma}} = ||(\mathbf{\Gamma}_{SS})^{-1}||_\infty$. $d$ is the maximum degree of the graph defined by $d = \max_i \sum_j \mathrm{I}(|\omega_{ij}| \neq 0)$ and $s$ is the number of true edges.

**Theorem** (Lemma 4, 5, 6, Ravikumar et al. (2011)). *Assume the irrepresentability condition holds with degree of* $\alpha \in (0, 1]$

$$\max_{e \in S^c} ||\mathbf{\Gamma}_{eS}\mathbf{\Gamma}_{SS}^{-1}||_1 \leq 1 - \alpha.$$

*If* $||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_{max} \leq \delta = \delta_{n,p}$ *and* $n$ *satisfies*

$$\delta_{n,p} \leq \left[6d(1 + 8\alpha^{-1}) \max\{\kappa_{\mathbf{\Gamma}^*}\kappa_{\mathbf{\Sigma}^*}, \kappa_{\mathbf{\Gamma}^*}^2\kappa_{\mathbf{\Sigma}^*}^3\}\right]^{-1},$$

*then we have*

1. $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_{max} \leq 2\kappa_{\mathbf{\Gamma}^*}\left(||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_{max} + 8\alpha^{-1}\delta\right) \leq 2\kappa_{\mathbf{\Gamma}^*}(1 + 8\alpha^{-1})\delta,$
2. $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_2 \leq 2\kappa_{\mathbf{\Gamma}^*}(1 + 8\alpha^{-1}) \min\{\sqrt{s + p}, d\}\delta,$
3. $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_F \leq 2\kappa_{\mathbf{\Gamma}^*}(1 + 8\alpha^{-1})\sqrt{s + p}\,\delta,$

*where* $\hat{\mathbf{\Omega}}$ *is the graphical lasso estimator that solves (4.1).*

We note that $\delta_{n,p}$ corresponds to $\bar{\delta}_f(n, p^\tau)$ in the original reference.

### CLIME

Let us introduce the class of a precision matrix used in Cai et al. (2011). For $0 \leq q < 1$,

$$\mathcal{U}(q, c_0(p)) = \left\{\mathbf{\Omega} \succ 0 : ||\mathbf{\Omega}||_1 \leq M, \max_{1 \leq i \leq p} \sum_{j=1}^{p} |\omega_{ij}|^q \leq s_0(p)\right\}.$$

**Theorem** (Theorem 6, Cai et al. (2011)). *If* $||\mathbf{\Omega}||_1||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_{max} \leq \delta$, *then we have*

1. $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_{max} \leq 4||\mathbf{\Omega}||_1 \delta$,

2. $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_2 \leq C s_0(p)(4||\mathbf{\Omega}||_1 \delta)^{1-q}$, *if* $\mathbf{\Omega} \in \mathcal{U}(q, c_0(p))$,

3. $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_F^2 / p \leq C s_0(p)(4||\mathbf{\Omega}||_1 \delta)^{2-q}$, *if* $\mathbf{\Omega} \in \mathcal{U}(q, c_0(p))$,

*where* $\hat{\mathbf{\Omega}}$ *is the CLIME estimator that solves (4.4) and* $C > 0$ *is a numerical constant.*

### Graphical Dantzig selector

The graphical Dantzig selector aims to solve $p$ optimization problems below (Yuan 2010)

$$\min_{\beta_j \in \mathbb{R}^{p-1}} ||\beta_j||_1, \quad \text{subject to } ||\hat{\mathbf{\Sigma}}^{plug}_{-j,j} - \hat{\mathbf{\Sigma}}^{plug}_{-j,-j}\beta_j||_\infty \leq \lambda, \tag{2.11}$$

for $j = 1, \ldots, p$. Let $d = \max_i \sum_j \mathrm{I}(|\omega_{ij}| \neq 0)$ (the maximum degree of the graph).

**Theorem** (A consequence of Lemma 11, Yuan (2010))**.** *Assume* $\mathbf{\Omega} \in O(v, \eta, \tau)$ *defined by*

$$O(v, \eta, \tau) = \Big\{ \mathbf{\Omega} \succ 0 : v^{-1} \leq \lambda_{min}(\mathbf{\Omega}) \leq \lambda_{max}(\mathbf{\Omega}) \leq v, ||\mathbf{\Sigma}\mathbf{\Omega} - \mathbf{I}||_{max} \leq \eta, ||\mathbf{\Omega}||_1 \leq \tau \Big\}.$$

*If* $\tau v ||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_{max} + \eta v \leq \delta$, *then we have*

$$||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_1 \leq C d \delta,$$

*where* $\hat{\mathbf{\Omega}}$ *is the graphical Dantzig estimator that solves (2.11) and* $C$ *depends only on* $v, \tau, \lambda_{min}(\mathbf{\Omega}), \lambda_{max}(\mathbf{\Omega})$.

Note that the $\ell_1$-norm of a matrix bounds the spectral norm, so we also have $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_2 \leq C d \delta$. □

# Chapter 3

# Relaxation of implicit assumptions

Estimation of the IPW estimator with missing data depends on two implicit assumptions other than Assumption 1, 2, and 3; known mean (or equivalently zero mean) and missing probability. In Chapter 3, we will relax such conditions and show corresponding concentration results.

## 3.1 The case of unknown mean

When the first moment of the underlying distribution is unknown, the IPW estimator should be modified accordingly, but the same rate $O_p(\sqrt{\log p/n})$ also holds true. We do not directly estimate the mean parameter $\mu_k$, but $\mu_k\mu_\ell$ because of the dependent missing structure.

Assume that we observe $\tilde{Y}_{ik} = \delta_{ik}\tilde{X}_{ik}$ where $\tilde{X}_{ik}$ has an unknown mean $\mu_k$. Adopting previous notations, we define $X_{ik}$ to satisfy $\tilde{X}_{ik} = X_{ik} + \mu_k$.

Then, it is easy to show that

$$\mathbb{E}\big[\sum_{i=1}^{n}\tilde{Y}_{ik}\tilde{Y}_{i\ell}\big] = n\pi_{k\ell}(\sigma_{k\ell}+\mu_k\mu_\ell), \quad \mathbb{E}\big[\sum_{i\neq j}^{n}\tilde{Y}_{ik}\tilde{Y}_{j\ell}\big] = n(n-1)\pi_k\pi_\ell\mu_k\mu_\ell.$$

With a simple calculation, we can define the unbiased covariance matrix estimator by $\hat{\boldsymbol{\Sigma}}^{IPW\mu} = \Big((\hat{\boldsymbol{\Sigma}}^{IPW\mu})_{k\ell}, 1\le k,\ell\le p\Big)$ with

$$(\hat{\boldsymbol{\Sigma}}^{IPW\mu})_{k\ell} = \frac{\sum_{i=1}^{n}\tilde{Y}_{ik}\tilde{Y}_{i\ell}}{n\pi_{k\ell}} - \frac{\sum_{i\neq j}^{n}\tilde{Y}_{ik}\tilde{Y}_{j\ell}}{n(n-1)\pi_k\pi_\ell}. \tag{3.1}$$

It is not difficult to find resemblance of (3.1) with the sample covariance matrix $\boldsymbol{S}$ when data is completely observed. The $(k,\ell)$-th component of $\boldsymbol{S}$ is defined by

$$\boldsymbol{S}_{k\ell} = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{X}_{ik}-\hat{\mu}_k)(\tilde{X}_{i\ell}-\hat{\mu}_\ell),$$

where $\hat{\mu}_k = n^{-1}\sum_{i=1}^{n}\tilde{X}_{ik}$, and it can be rearranged by

$$\boldsymbol{S}_{k\ell} = \frac{\sum_{i=1}^{n}\tilde{X}_{ik}\tilde{X}_{i\ell}}{n} - \frac{\sum_{i\neq j}^{n}\tilde{X}_{ik}\tilde{X}_{j\ell}}{n(n-1)},$$

which is equal to (3.1) when $\pi_{k\ell}=\pi_k=1$ for all $k,\ell$.

Let $(k,\ell)$ be a dual in $\{1,\dots,p\}^2$. Using $\tilde{Y}_{ik} = \delta_{ik}X_{ik} + \delta_{ik}\mu_k = Y_{ik} + \delta_{ik}\mu_k$, we can decompose the first term in (3.1) as follows.

$$\begin{aligned}
&\frac{\sum_{i=1}^{n}\tilde{Y}_{ik}\tilde{Y}_{i\ell}}{n\pi_{k\ell}} - (\sigma_{k\ell}+\mu_k\mu_\ell) \\
= &\left\{\frac{\sum_{i=1}^{n}Y_{ik}Y_{i\ell}}{n\pi_{k\ell}} - \sigma_{k\ell}\right\} + \left\{\frac{\sum_{i=1}^{n}\delta_{ik}\delta_{i\ell}\mu_k X_{i\ell}}{n\pi_{k\ell}}\right\} \\
&+\left\{\frac{\sum_{i=1}^{n}\delta_{ik}\delta_{i\ell}X_{ik}\mu_\ell}{n\pi_{k\ell}}\right\} + \left\{\frac{\sum_{i=1}^{n}\delta_{ik}\delta_{i\ell}\mu_k\mu_\ell}{n\pi_{k\ell}} - \mu_k\mu_\ell\right\} \\
= &\ A_1 + A_2 + A_3 + A_4.
\end{aligned}$$

A deviation inequality for $A_1$ comes from Lemma 4. On the other hands, since $A_2$, $A_3$, and $A_4$ are independent sum of sub-Gaussian variables, the related concentration inequalities are already well-established, which, in our case, can be found in Lemma 1 (a) and 8. The second term in (3.1) can be decomposed by

$$
\begin{aligned}
& \frac{\sum_{i\neq j}\tilde{Y}_{ik}\tilde{Y}_{j\ell}}{n(n-1)\pi_k\pi_\ell} - \mu_k\mu_\ell \\
= \; & \frac{\sum_{i\neq j}(\tilde{Y}_{ik}-\mathbb{E}\tilde{Y}_{ik})(\tilde{Y}_{j\ell}-\mathbb{E}\tilde{Y}_{j\ell})}{n(n-1)\pi_k\pi_\ell} + \frac{\sum_{i\neq j}(\tilde{Y}_{ik}-\mathbb{E}\tilde{Y}_{ik})\mathbb{E}\tilde{Y}_{j\ell}}{n(n-1)\pi_k\pi_\ell} + \frac{\sum_{i\neq j}(\tilde{Y}_{i\ell}-\mathbb{E}\tilde{Y}_{i\ell})\mathbb{E}\tilde{Y}_{ik}}{n(n-1)\pi_k\pi_\ell} \\
= \; & \frac{\sum_{i\neq j}(\tilde{Y}_{ik}-\mathbb{E}\tilde{Y}_{ik})(\tilde{Y}_{j\ell}-\mathbb{E}\tilde{Y}_{j\ell})}{n(n-1)\pi_k\pi_\ell} + \frac{\mu_\ell\sum_{i=1}^n(\tilde{Y}_{ik}-\mathbb{E}\tilde{Y}_{ik})}{n\pi_k} + \frac{\mu_k\sum_{i=1}^n(\tilde{Y}_{i\ell}-\mathbb{E}\tilde{Y}_{i\ell})}{n\pi_\ell} \\
= \; & \frac{\sum_{i\neq j}(\tilde{Y}_{ik}-\mathbb{E}\tilde{Y}_{ik})(\tilde{Y}_{j\ell}-\mathbb{E}\tilde{Y}_{j\ell})}{n(n-1)\pi_k\pi_\ell} + \frac{\mu_\ell\sum_{i=1}^n\delta_{ik}X_{ik}}{n\pi_k} + \frac{\mu_\ell\sum_{i=1}^n(\delta_{ik}-\pi_k)}{n\pi_k} \\
& + \frac{\mu_k\sum_{i=1}^n\delta_{i\ell}X_{i\ell}}{n\pi_\ell} + \frac{\mu_k\sum_{i=1}^n(\delta_{i\ell}-\pi_\ell)}{n\pi_\ell} \\
= \; & B_1+B_2+B_3+B_4+B_5.
\end{aligned}
$$

The concentration of each term except $B_1$ is easily derived using Lemma 1 (a) and 8. To analyze the concentration of $B_1$ which is a dependent sum of cross-product of sub-Gaussian variables, we need a new version of Hanson-Wright inequality. Lemma 5 is more general than that given in Rudelson and Vershynin (2013) in the sense that two random variables $X_i, Y_i$ are not necessarily equal. The generalization is possible because of the decoupling technique from which we can separately handle $\{X_i : i \in \Lambda\}$ and $\{Y_i : i \notin \Lambda\}$

for some $\Lambda \subset \{1, \ldots, n\}$.

**Lemma 5.** *Let $(X, Y)$ be a pair of (possibly correlated) random variables satisfying $\mathbb{E}X = \mathbb{E}Y = 0$, and*

$$\sup_{r \geq 1} \frac{\{\mathbb{E}|X|^r\}^{1/r}}{\sqrt{r}} \leq K_X, \quad \sup_{r \geq 1} \frac{\{\mathbb{E}|Y|^r\}^{1/r}}{\sqrt{r}} \leq K_Y.$$

*Assume $n$ copies $\{(X_i, Y_i)\}_{i=1}^n$ of $(X, Y)$ are independently observed. For a matrix $\boldsymbol{A} = (a_{ij}, 1 \leq i, j \leq n)$ with zero diagonals $a_{ii} = 0$, we have that*

$$\mathrm{P}\Big[|\sum_{i \neq j} a_{ij} X_i Y_j| > t\Big] \leq 2 \exp\Big\{-c \min\Big(\frac{t^2}{K_X^2 K_Y^2 ||\boldsymbol{A}||_F^2}, \frac{t}{K_X K_Y ||\boldsymbol{A}||_2}\Big)\Big\}, \quad t \geq 0.$$

*for some numerical constant $c > 0$.*

*Proof.* Without loss of generality, we assume $K_X = K_Y = 1$. Let $\{\eta_i\}_{i=1}^n$ be independent Bernoulli variables with success probability $1/2$. Then, by observing $\mathbb{E}\eta_i(1-\eta_j) = \mathrm{I}(i \neq j)/4$, it can be seen that $S \equiv \sum_{i \neq j} a_{ij} X_i Y_j = 4\mathbb{E}_{\{\eta_i\}} S_\eta$ where $S_\eta = \sum_{i,j} \eta_i(1-\eta_j) a_{ij} X_i Y_j$ and $\mathbb{E}_{\{\eta_i\}}$ is an expectation taken over $\{\eta_i\}$. Let $\Lambda_\eta = \{i : \eta_i = 1\}$ be the index set of successes. Since $S_\eta = \sum_{i \in \Lambda_\eta, j \in \Lambda_\eta^c} a_{ij} X_i Y_j$ is a function of $\{Y_j : j \in \Lambda_\eta^c\}$ given $\{\eta_i\}$ and $\{X_i : i \in \Lambda_\eta\}$, $S_\eta$ conditionally follows is a sub-Gaussian distribution.

We assume $\{\eta_i\}$ is conditioned on all the following statements unless specified otherwise. Then, the previous results yield

$$\begin{aligned} \mathbb{E}_{\{(X_j, Y_j) : j \in \Lambda_\eta^c\}}\Big[\exp(4\lambda S_\eta)\Big|\{X_i : i \in \Lambda_\eta\}\Big] &= \mathbb{E}_{\{Y_j : j \in \Lambda_\eta^c\}}\Big[\exp(4\lambda S_\eta)\Big|\{X_i : i \in \Lambda_\eta\}\Big] \\ &\leq \exp\Big\{c\lambda^2 \textstyle\sum_{j \in \Lambda_\eta^c} (\sum_{i \in \Lambda_\eta} a_{ij} X_i)^2\Big\}, \end{aligned}$$

where the equality holds since $\exp(4\lambda S_\eta)$ does not depend on $\{X_j\}_{j \in \Lambda_\eta^c}$ and the inequality is from sub-Gaussianity of $S_\eta$. Taking expectation with

respect to $\{X_i : i \in \Lambda_\eta\}$ on both sides, we get the following result;

$$\begin{aligned}\mathbb{E}_{\{X_i:i\in\Lambda_\eta\},\{(X_j,Y_j):j\in\Lambda_\eta^c\}}\big[\exp(4\lambda S_\eta)\big] &\leq \mathbb{E}_{\{X_i:i\in\Lambda_\eta\}}\Big[\exp\Big\{c\lambda^2\textstyle\sum_{j\in\Lambda_\eta^c}(\sum_{i\in\Lambda_\eta}a_{ij}X_i)^2\Big\}\Big] \\ &= \mathbb{E}_{\{X_i\}}\Big[\exp\Big\{c\lambda^2\textstyle\sum_{j\in\Lambda_\eta^c}(\sum_{i\in\Lambda_\eta}a_{ij}X_i)^2\Big\}\Big],\end{aligned}$$

where the equality holds from independence among $n$ samples. Also, since the left-hand side does not depend on $\{Y_i : i \in \Lambda_\eta\}$, we get

$$\mathbb{E}_{\{(X_i,Y_i)\}}\big[\exp(4\lambda S_\eta)\big|\{\eta_i\}\big] \leq \mathbb{E}_{\{X_i\}}\Big[\exp\Big\{c\lambda^2\sum_{j\in\Lambda_\eta^c}\big(\sum_{i\in\Lambda_\eta}a_{ij}X_i\big)^2\Big\}\Big|\{\eta_i\}\Big](\equiv T_\eta),$$

where we begin to display the conditional dependency on $\{\eta_i\}$. Following the step 3 and 4 in Rudelson and Vershynin (2013), we can achieve an uniform bound of $T_\eta$ independent of $\{\eta_i\}$ and thus get

$$T_\eta \leq \exp\{C\lambda^2||A||_F^2\} \quad \text{for } \lambda \leq c/||A||_2,$$

for some positive constants $c$ and $C$. Then, we have

$$\begin{aligned}\mathbb{E}\big[\exp(\lambda S)\big] &= \mathbb{E}\big[\exp(\mathbb{E}_{\{\eta_i\}}4\lambda S_\eta)\big] \\ &\leq \mathbb{E}_{\{(X_i,Y_i)\}_i,\{\eta_i\}}\big[\exp(4\lambda S_\eta)\big](\because \text{Jensen's inequality}) \\ &= \mathbb{E}\Big[\mathbb{E}\big[\exp(4\lambda S_\eta)\big|\{\eta_i\}\big]\Big] \\ &= \mathbb{E}[T_\eta] \\ &\leq \exp\{C\lambda^2||A||_F^2\} \quad \text{for } \lambda \leq c/||A||_2,\end{aligned}$$

Following the step 5 in Rudelson and Vershynin (2013), we can get the concentration of $S$ given in the lemma below. Let $||X||_{\psi_2}$ be a $\psi_2$-norm of $X$ defined by

$$||X||_{\psi_2} = \inf\Big\{R > 0 : \mathbb{E}e^{\frac{|X|^2}{R^2}} \leq 2\Big\}.$$

**Lemma.** *Let $(X, Y)$ be a pair of (possibly correlated) random variables satisfying $\mathbb{E}X = \mathbb{E}Y = 0$, and*

$$||X||_{\psi_2} \leq K_X, ||Y||_{\psi_2} \leq K_Y. \tag{3.2}$$

*Assume $n$ samples $\{(X_i, Y_i)\}_{i=1}^n$ are identically and independently observed. For a matrix $A = (a_{ij}, 1 \leq i, j \leq n)$ with zero diagonals, we have that*

$$\mathrm{P}\Big[\Big|\sum_{i\neq j} a_{ij}X_iY_j\Big| > t\Big] \leq 2\exp\Big\{-c\min\Big(\frac{t^2}{K_X^2K_Y^2||A||_F^2}, \frac{t}{K_XK_Y||A||_2}\Big)\Big\}, \quad t \geq 0.$$

*for some numerical constant $c > 0$.*

Note that the finite $\psi_2$-norm in (3.2) characterizes a sub-Gaussian random variable and can be replaced by the uniformly bounded moments in (2.6), since $\sup_{r\geq 1}\{\mathbb{E}|X|^r\}^{1/r}/\sqrt{r} \leq K$ implies $||X||_{\psi_2} \leq 2eK$. In other words, provided $X_i$ and $Y_j$ satisfy the moment condition with constants $K_X$ and $K_Y$, respectively, the conclusion of the lemma above still holds (with different $c$). This completes the proof. □

Now, we get the concentration bound for $B_1$ using the lemma above;

$$\mathrm{P}\Big[\Big|\frac{\sum_{i\neq j}(\tilde{Y}_{ik} - \mathbb{E}\tilde{Y}_{ik})(\tilde{Y}_{j\ell} - \mathbb{E}\tilde{Y}_{j\ell})}{n(n-1)\pi_k\pi_\ell}\Big| > t\Big] \leq 2\exp\Big\{-\frac{c\pi_k\pi_\ell nt}{\sigma_{kk}^{1/2}\sigma_{\ell\ell}^{1/2}K^2}\Big\},$$

for $t \geq \dfrac{\sigma_{kk}^{1/2}\sigma_{\ell\ell}^{1/2}K^2}{\pi_k\pi_\ell n}$, since the matrix in $\mathbb{R}^{n\times n}$ with off-diagonals 1 and diagonals 0 has both Frobenius and spectral norms of being $n-1$. By

plugging-in $t \leftarrow \frac{t\sigma_{kk}^{1/2}\sigma_{\ell\ell}^{1/2}K^2}{c\pi_k\pi_\ell}\sqrt{\frac{\log p}{n}}$, we have

$$\mathrm{P}\left[\left|\frac{\sum_{i\neq j}(\tilde{Y}_{ik}-\mathbb{E}\tilde{Y}_{ik})(\tilde{Y}_{j\ell}-\mathbb{E}\tilde{Y}_{j\ell})}{n(n-1)\pi_k\pi_\ell}\right| > \frac{t\sigma_{kk}^{1/2}\sigma_{\ell\ell}^{1/2}K^2}{\pi_k\pi_\ell}\sqrt{\frac{\log p}{n}}\right] \;\leq\; 2\exp\left\{-t\sqrt{n\log p}\right\}$$

for $t\sqrt{n\log p} \geq c$. Then, if we assume $n > \log p$, the probability above is bounded by $2p^{-t}$.

Combining all results for $A_1, \ldots, A_4$, $B_1, \ldots, B_5$, we can derive the concentration inequality for each component of $\hat{\mathbf{\Sigma}}^{IPW\mu}$, and thus the following theorem.

**Theorem 2.** *Assume the conditions of Lemma 4 hold except a mean zero condition, and further assume the sample size and dimension satisfy*

$$n/\log p > c\max\left\{\frac{1}{R\big(2K/\sqrt{v_{min}}\big)},\ \frac{K^2}{\pi_{min,d}+2e^2K^2}\right\},$$

*then it holds that*

$$\mathrm{P}\left[\left|\left|\hat{\mathbf{\Sigma}}^{IPW\mu}-\mathbf{\Sigma}\right|\right|_{max} \geq C\sqrt{\frac{\log p}{n}}\right] \leq dp^{-1},$$

*where $c > 0, d > 0$ are numerical constants and $C > 0$ is a constant depending only on $K$, $\sigma_{max}$, $\max_k|\mu_k|$, $\pi_{min}$, and $\min_k \pi_k$.*

Proof of Theorem 2 is very similar to that of Theorem 1, so we do not provide the details.

**Remark.** *In the theorem above, dependency of the constant $C$ on the parameters can be specified as, (up to a constant factor)*

$$C = \frac{\max\{\sigma_{max}, \mu_{max}, \mu_{max}^2\}\max\left\{K^2\sqrt{R(K)}, \sqrt{1+2e^2K^2}\right\}}{\min\{\pi_{min}^{3/2}, \pi_{min,d}^2\}}$$

*where* $\mu_{max} = \max_k |\mu_k|$ *and* $\pi_{min,d} = \min_k \pi_k$. *Supposedly, dependency on the mean parameter* $\mu_{max}$ *can be taken away in* $C$ *if a missing value is filled by the empirical mean of available data. However, we leave this as the future work.*

## 3.2 The case of unknown missing probability

In real applications, the missing probability $\pi_{jk}$ is rarely known, but to be estimated. Let $\hat{\pi}_{jk}$ be any estimate satisfying $\hat{\pi}_{jk} > 0, \forall j, k$, with high probability. Then, the resulting IPW estimator is presented by

$$\hat{\mathbf{\Sigma}}^{IPW\pi} = \Big( (\hat{\mathbf{\Sigma}}^{IPW})_{jk} \frac{\pi_{jk}}{\hat{\pi}_{jk}},\ 1 \leq j, k \leq p \Big), \tag{3.3}$$

provided that the population mean is known for the sake of simplicity. The following lemma shows how the concentration of (3.3) is related to that of $\hat{\pi}_{jk}$.

**Lemma 6.** *Assume*

$$\max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}| < B_1, \quad \hat{\pi}_{k\ell} > 0, \forall k, \ell,$$

$$\big|\big| \boldsymbol{S}_Y - \mathbf{\Sigma}^{\pi} \big|\big|_{max} < B_2$$

$$\big|\big| \hat{\mathbf{\Sigma}}^{IPW} - \mathbf{\Sigma} \big|\big|_{max} < B_3$$

*where* $B_1, B_2$, *and* $B_3$ *are positive constants. Then, we have*

$$\big|\big| \hat{\mathbf{\Sigma}}^{IPW\pi} - \mathbf{\Sigma} \big|\big|_{max} \leq B_1 B_2 + B_1 \sigma_{max} + B_3.$$

*Proof.* By the triangular inequality, we observe

$$\begin{aligned}
\big|\big|\hat{\boldsymbol{\Sigma}}^{IPW\pi} - \boldsymbol{\Sigma}\big|\big|_{max} &\leq \big|\big|\hat{\boldsymbol{\Sigma}}^{IPW\pi} - \hat{\boldsymbol{\Sigma}}^{IPW}\big|\big|_{max} + \big|\big|\hat{\boldsymbol{\Sigma}}^{IPW} - \boldsymbol{\Sigma}\big|\big|_{max} \\
&\leq \max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}| \cdot \big|\big|\boldsymbol{S}_Y\big|\big|_{max} + \big|\big|\hat{\boldsymbol{\Sigma}}^{IPW} - \boldsymbol{\Sigma}\big|\big|_{max} \\
&\leq \max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}| \cdot \big|\big|\boldsymbol{S}_Y - \boldsymbol{\Sigma}^{\pi}\big|\big|_{max} \\
&\quad + \max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}| \cdot \big|\big|\boldsymbol{\Sigma}^{\pi}\big|\big|_{max} + \big|\big|\hat{\boldsymbol{\Sigma}}^{IPW} - \boldsymbol{\Sigma}\big|\big|_{max}
\end{aligned}$$

where $\boldsymbol{S}_Y = n^{-1}\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}$ and $\boldsymbol{\Sigma}^{\pi} = \Big(\pi_{jk}\sigma_{jk}, 1 \leq j,k \leq p\Big)$. Thus, we get

$$\big|\big|\hat{\boldsymbol{\Sigma}}^{IPW\pi} - \boldsymbol{\Sigma}\big|\big|_{max} \leq B_1 B_2 + B_1\big|\big|\boldsymbol{\Sigma}^{\pi}\big|\big|_{max} + B_3.$$

Finally, we note that

$$\big|\big|\boldsymbol{\Sigma}^{\pi}\big|\big|_{max} \leq \big|\big|\boldsymbol{\Sigma}\big|\big|_{max} = \sigma_{max}$$

where the last equality holds for a symmetric positive definite matrix. □

When an additional information on missing is not available, it is natural to use the empirical proportions $\hat{\pi}_{jk}^{emp} = n^{-1}\sum_{i=1}^n \delta_{ij}\delta_{ik}$ of observed samples for estimation of $\pi_{jk}$ since it is asymptotically unbiased for $\pi_{jk}$ (by the law of large numbers). Lemma 7 describes the concentration of its inverse probability.

**Lemma 7.** *Assume the sample size and dimension satisfy $n/\log p > C/\pi_{min}$ for some numerical constant $C > 0$. Then, it holds that with probability at*

*most* $2/p$

$$\max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}^{emp}| \geq \sqrt{\frac{C \log p}{\pi_{min}^2 n}}, \ \textit{and} \ \hat{\pi}_{k\ell}^{emp} > 0, \forall k, \ell. \tag{3.4}$$

*Proof.* First, we observe that on the event $G = G_{n,p} = \{\hat{\pi}_{k\ell}^{emp} > 0, \forall k, \ell\}$, we have for $t > 0$

$$|1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}^{emp}| \geq t \quad \Leftrightarrow \quad (1 - t\pi_{k\ell})\hat{\pi}_{k\ell}^{emp} \geq \pi_{k\ell} \text{ or } (1 + t\pi_{k\ell})\hat{\pi}_{k\ell}^{emp} \leq \pi_{k\ell}.$$

Let $A_{k\ell} = \{(1 - t\pi_{k\ell})\hat{\pi}_{k\ell}^{emp} \geq \pi_{k\ell}\}$ and $B_{k\ell} = \{(1 + t\pi_{k\ell})\hat{\pi}_{k\ell}^{emp} \leq \pi_{k\ell}\}$. Using these notations, we get

$$\begin{aligned} \mathrm{P}\Big[\Big\{ \max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}^{emp}| \geq t \Big\} \cap G\Big] &= \mathrm{P}\Big[G \cap \big\{ \cup_{k,\ell} (A_{k\ell} \cup B_{k\ell}) \big\}\Big] \\ &\leq \mathrm{P}\big[ \cup_{k,\ell} (A_{k\ell} \cup B_{k\ell})\big] \\ &\leq \sum_{k,\ell} \mathrm{P}(A_{k\ell} \cup B_{k\ell}). \end{aligned}$$

We introduce the deviation inequality for a sum of Bernoulli variables.

**Lemma 8** (Boucheron et al. (2016), p 48)**.** *Let* $\{\delta_i\}_{i=1}^n$ *be independent Bernoulli variables with probability* $\pi$ *of being* 1*. Then, there exists a numerical constant* $C > 0$ *such that for* $t > 0$*,*

$$\mathrm{P}\Big[ \pm \sum_{i=1}^{n} (\delta_i - \pi) \geq nt \Big] \leq \exp(-Cn\pi t^2).$$

If $t < \pi_{k\ell}^{-1}$, by using Lemma 8, it holds

$$\mathrm{P}(A_{k\ell}) = \mathrm{P}\Big[\hat{\pi}_{k\ell}^{emp} - \pi_{k\ell} \geq \frac{t\pi_{k\ell}}{1 - t\pi_{k\ell}}\Big] \leq \exp\Big\{ - \frac{Cnt^2\pi_{k\ell}^3}{(1 - t\pi_{k\ell})^2} \Big\}.$$

Similarly, we have

$$\mathrm{P}(B_{k\ell}) \leq \exp\Big\{ - \frac{Cnt^2\pi_{k\ell}^3}{(1+t\pi_{k\ell})^2}\Big\}.$$

If we define $\pi_{min} = \min_{k,\ell} \pi_{k\ell}$, we get by the union argument

$$\begin{aligned} \mathrm{P}(A_{k\ell} \cup B_{k\ell}) &\leq 2\exp\Big\{ - \frac{Cnt^2\pi_{k\ell}^3}{(1+t\pi_{k\ell})^2}\Big\} \\ &\leq 2\exp\Big\{ - \frac{Cnt^2\pi_{min}^3}{(1+t\pi_{min})^2}\Big\}, \end{aligned}$$

where the last inequality depends on monotonicity of $x \in (0,1) \mapsto \dfrac{x^3}{(1+tx)^2}$ for $t > 0$. Combining these results, we can conclude

$$\mathrm{P}\Big[\Big\{ \max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}^{emp}| \geq t\Big\} \cap G\Big] \leq 2p^2 \exp\Big\{ - \frac{Cnt^2\pi_{min}^3}{(1+t\pi_{min})^2}\Big\}.$$

If $t \leftarrow \sqrt{4\log p/(C\pi_{min}^2 n)}$ and assume $n/\log p > 12/(C\pi_{min})$, then we can derive

$$\mathrm{P}\Big[\Big\{ \max_{k,\ell} |1/\pi_{k\ell} - 1/\hat{\pi}_{k\ell}^{emp}| \geq \frac{2}{\pi_{min}}\sqrt{\frac{\log p}{Cn}}\Big\} \cap G\Big] \leq \frac{2}{p},$$

which completes the proof. □

We denote the empirical version $\hat{\mathbf{\Sigma}}^{emp}$ of the IPW estimator by

$$(\hat{\mathbf{\Sigma}}^{emp})_{jk} = \frac{\sum_{i=1}^{n} \delta_{ij}\delta_{ik}X_{ij}X_{ik}}{\sum_{i=1}^{n} \delta_{ij}\delta_{ik}}, \quad 1 \leq j,k \leq p, \tag{3.5}$$

which corresponds to (3.3) with $\hat{\pi}_{jk}^{emp}$ in place of $\hat{\pi}_{jk}$. One may realize the equivalence of the empirical estimate (3.5) to a pairwise complete analysis.

**Theorem 3.** *Assume the conditions of Lemma 4 without knowing missing probabilities, and further assume the sample size and dimension satisfy*

$$n/\log p > c \max\left\{ \frac{1}{R\big(2K/\sqrt{v_{min}}\big)},\ \frac{1}{\pi_{min}} \right\},$$

*then it holds that*

$$\mathrm{P}\left[ \left|\left| \hat{\mathbf{\Sigma}}^{emp} - \mathbf{\Sigma} \right|\right|_{max} \geq C\sqrt{\frac{\log p}{n}} \right] \leq dp^{-1},$$

*where $c > 0, d > 0$ are numerical constants and $C > 0$ is a constant depending only on $K$, $\sigma_{max}$, and $\pi_{min}$.*

Theorem 3 is not difficult to show if Lemma 4, 6, and 7 are used together. This result has an implication that the convergence rate $\sqrt{\log p/n}$ in Theorem 1 is preserved, and thus the same statements in Theorem 2 hold true with $\hat{\mathbf{\Sigma}}^{emp}$. It should be pointed out that Kolar and Xing (2012) use the estimator $\hat{\mathbf{\Sigma}}^{emp}$, while their theory is limited to the independent missing structure. Thus, Theorem 3 justifies their theory for the empirical IPW estimator even under the dependent structure.

**Remark.** *In the theorem above, dependency of the constant $C$ on the parameters can be specified as, (up to a constant factor)*

$$C = \frac{\sigma_{max} \max\left\{K^2\sqrt{R(K)}, 1\right\}}{\pi_{min}}.$$

# Chapter 4

# Non-positive semi-definiteness of the plug-in estimator

Despite its straightforward derivation and applicability to multivariate procedures in the presence of missing, the IPW estimator has one critical issue in a practical point of view; non-positive semi-definiteness (non-PSDness). Note that this does not cause problems in the convergence rate, since the norm is element-wisely defined. It is well known that the element-wise product of two matrices may not preserve a nice property of the matrices. As addressed in high-dimensional covariance estimation (thresholding, banding, and tapering) (Bickel and Levina 2008a; Rothman et al. 2009), the positive semi-definiteness is one of the typical examples to be broken down by the Hadamard product of a positive semi-definite (PSD) matrix and a general matrix. This is also the case for the IPW estimator, which makes

it practically difficult to use the IPW estimator when using implemented algorithms for a precision matrix. For instance, we can plug-in the IPW estimator into the graphical lasso or the CLIME to estimate sparse precision matrix $\mathbf{\Omega} = (\omega_{k\ell}, 1 \leq k, \ell \leq p)$, when missing data is available. However, the popularly used algorithms (`glasso` package or `clime` package in R) require the plugged-in estimator to be positive semi-definite. In Chapter 4, we examine their algorithms from this point of view and also suggest possible solutions.

In what follows, we differentiate a plug-in matrix (estimator) $\hat{\mathbf{\Sigma}}^{plug}$ and an initial matrix (estimator) $\mathbf{\Sigma}^{(0)}$ (or $\mathbf{\Omega}^{(0)}$) that is used to initialize iterative steps.

## 4.1 Graphical lasso

The graphical lasso proposed by Friedman et al. (2008) aims to maximize the penalized likelihood function

$$\max_{\mathbf{\Omega} \succeq 0} \Big\{ \log |\mathbf{\Omega}| - \mathrm{tr}(\mathbf{\Omega}\hat{\mathbf{\Sigma}}^{plug}) - \lambda \sum_{k,\ell} |\omega_{k\ell}| \Big\}, \tag{4.1}$$

for a penalty parameter $\lambda > 0$. To solve (4.1), a coordinate descent algorithm described in Algorithm 1 is proposed by Friedman et al. (2008) and implemented in R package `glasso`.

**Algorithm 1** The coordinate descent algorithm for the graphical lasso

**Input:** An initial matrix $\mathbf{\Sigma}^{(0)}$ of $\mathbf{\Sigma}$, the plug-in matrix $\hat{\mathbf{\Sigma}}^{plug}$

1: **for** $i = 1, 2, \ldots,$ **do**

2: **for** $j = 1, \ldots, p,$ **do**

3: Solve the least squared regression with the $\ell_1$-penalty

$$\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}^{p-1}} \frac{1}{2} \beta^{\mathrm{T}} \mathbf{\Sigma}^{(i-1)}_{\setminus j \setminus j} \beta - \beta^{\mathrm{T}} \hat{\mathbf{\Sigma}}^{plug}_j + \lambda ||\beta||_1, \tag{4.2}$$

where $\mathbf{\Sigma}^{(i-1)}_{\setminus j \setminus j}$ is obtained by removing the $j$-th row and column in $\mathbf{\Sigma}^{(i-1)}$ and $\hat{\mathbf{\Sigma}}^{plug}_j$ is the $j$-th column of $\hat{\mathbf{\Sigma}}^{plug}$ without the $j$-th entry.

4: Replace the $j$-th column and row of off-diagonal entries in $\mathbf{\Sigma}^{(i-1)}$ with $\mathbf{\Sigma}^{(i-1)}_{\setminus j \setminus j} \hat{\beta}_j$.

5: **end for**

6: Let $\mathbf{\Sigma}^{(i)} \leftarrow \mathbf{\Sigma}^{(i-1)}$.

7: **end for**

8: Let $\mathbf{\Sigma}^{(\infty)}$ and $\{\hat{\beta}_1, \ldots, \hat{\beta}_p\}$ be the final outputs from lines 1-7.

9: **for** $j = 1, \ldots, p$ **do**

10: $\hat{\mathbf{\Omega}}_{jj} = \left(\mathbf{\Sigma}^{(\infty)}_{jj} - \hat{\beta}_j^{\mathrm{T}} \mathbf{\Sigma}^{(\infty)}_j\right)^{-1}$ and $\hat{\mathbf{\Omega}}_j = -\hat{\mathbf{\Omega}}_{jj} \hat{\beta}_j$.

11: **end for**

**Output:** $\hat{\mathbf{\Omega}}$: the final estimate.

One can easily see that the optimization problem (4.1) is convex regardless of $\hat{\mathbf{\Sigma}}^{plug}$ ($\because$ the trace term is a linear function in $\mathbf{\Omega}$), but PSDness of $\hat{\mathbf{\Sigma}}^{plug}$ is needed when the algorithm is initialized.

First, PDness of $\mathbf{\Sigma}^{(i-1)}$ is required in (4.2) to find a well-defined solution of the lasso problem. Since PD $\mathbf{\Sigma}^{(i-1)}$ guarantees the updated matrix $\mathbf{\Sigma}^{(i)}$ to

be PD (Banerjee et al. 2008), the PD initial $\mathbf{\Sigma}^{(0)}$ is necessary to make sure every step runs successfully. However, currently available R packages (e.g. `glasso` version 1.10 from Friedman et al. (2008) or `huge` version 1.3.2 from Zhao et al. (2012)) set $\mathbf{\Sigma}^{(0)} \leftarrow \hat{\mathbf{\Sigma}}^{plug} + \lambda \mathbf{I}$ where $\lambda$ is the same parameter used in (4.1). As a consequence, unless $\lambda$ is bigger than the absolute value of the smallest (possibly negative) eigenvalue of $\hat{\mathbf{\Sigma}}^{IPW}$, the coordinate descent algorithm would fail to converge. For this reason, we propose to use the following inputs

$$\hat{\mathbf{\Sigma}}^{plug} \leftarrow \hat{\mathbf{\Sigma}}^{IPW}, \quad \mathbf{\Sigma}^{(0)} \leftarrow \mathrm{diag}\big(\hat{\mathbf{\Sigma}}^{IPW} + \lambda \mathbf{I}\big). \tag{4.3}$$

The choice for the initial matrix is because diagonals of the solution $\mathbf{\Sigma}^{(\infty)}$ should satisfy

$$\mathbf{\Sigma}^{(\infty)}_{ii} = \hat{\mathbf{\Sigma}}^{plug}_{ii} + \lambda, \forall i,$$

by the subgradient condition of (4.1), as noted in Friedman et al. (2008), and because diagonals of $\mathbf{\Sigma}^{(i)}$ do not change as iterations proceed. To use these proposed inputs, one should modify the off-the-shelf codes (e.g. `glasso` function in `glasso` package) since they do not currently allow users to control $\mathbf{\Sigma}^{(0)}$ and $\hat{\mathbf{\Sigma}}^{plug}$ individually.

Last but not least, it should be remarked that there is an algorithm developed to solve (4.1) by approximating the Hessian function (R package `QUIC` from Hsieh et al. (2014)). This method does not suffer from the PSDness issue discussed here (see Chapter 5.5). However, solving a similar difficulty in the other multivariate procedures remains open.

## 4.2 CLIME

We analyze the CLIME method proposed by Cai et al. (2011), which solves

$$\min_{\mathbf{\Omega}\in\mathbb{R}^{p\times p}} |\mathbf{\Omega}|_1 \quad \text{s.t.} \quad ||\hat{\mathbf{\Sigma}}^{plug}\mathbf{\Omega} - \mathbf{I}||_{max} \leq \lambda. \tag{4.4}$$

Cai et al. (2011) divide (4.4) into $p$ column-wise problems and relax the individual problem to be a linear programming, which leads to Algorithm 2.

**Algorithm 2** The CLIME algorithm

**Input:** An initial matrix $\mathbf{\Omega}^{(0)}$ of $\mathbf{\Omega}$, the plug-in matrix $\hat{\mathbf{\Sigma}}^{plug}$.

1: **for** $j = 1, \ldots, p$, **do**

2: Solve the linear programming below. We use the $j$-th column of $\mathbf{\Omega}^{(0)}$ for initialization of $\beta_j$

$$\begin{gathered}(\hat{r}, \hat{\beta}_j) = \arg\min_{r,\beta_j\in\mathbb{R}^p} ||r||_1 \\ \text{subject to } |\beta_j| \leq r(\text{element-wise}), ||\hat{\mathbf{\Sigma}}^{plug}\beta_j - e_j||_{max} \leq \lambda.\end{gathered} \tag{4.5}$$

3: **end for**

**Output:** $\hat{\mathbf{\Omega}} = [\hat{\beta}_1, \ldots, \hat{\beta}_p]$: the final estimate.

It is easily seen that the optimization problem (4.4) is convex regardless of the plug-in matrix. Moreover, Algorithm 2 does not require any constraint in the two inputs for a well-defined solution, contrary to Algorithm 1. However, the current implementations (e.g. `clime` version 0.4.1 from Cai et al. (2011), `fastclime` version 1.4.1 from Pang et al. (2014)) set the initial by solving $\mathbf{\Omega}^{(0)}(\hat{\mathbf{\Sigma}}^{plug} + \lambda\mathbf{I}) = \mathbf{I}$, which is not applicable to our case since an

initialization from $\boldsymbol{\Omega}^{(0)}(\hat{\boldsymbol{\Sigma}}^{IPW} + \lambda \mathbf{I}) = \mathbf{I}$ is not well-posed unless $\hat{\boldsymbol{\Sigma}}^{IPW} + \lambda \mathbf{I}$ is positive definite. Katayama et al. (2018) also point out that the solution of (4.4) may not exist, unless an input matrix $\hat{\boldsymbol{\Sigma}}^{plug}$ is guaranteed to be PSD. We conjecture this irregularity is due to the initialization. Thus, our proposal for the inputs is

$$\hat{\boldsymbol{\Sigma}}^{plug} \leftarrow \hat{\boldsymbol{\Sigma}}^{IPW}, \quad \boldsymbol{\Omega}^{(0)} \leftarrow \mathrm{diag}(\hat{\boldsymbol{\Sigma}}^{IPW})^{-1}.$$

Similarly to the graphical lasso, one should modify the implemented R functions (e.g. `clime` in `clime` package) to separately handle two inputs, since it is not allowed for now to control two input matrices $\boldsymbol{\Omega}^{(0)}$ and $\hat{\boldsymbol{\Sigma}}^{plug}$ independently.

## 4.3 More general solution: matrix approximation

Previously, we present the solutions that are specific to the precision matrix estimation problem, but we can circumvent the non-PSD issue for general statistical procedures. The idea is to approximate $\hat{\boldsymbol{\Sigma}}^{plug}$ by the nearest PSD matrix, which can be achieved by

$$\hat{\boldsymbol{\Sigma}}^{psd} = \arg\min_{\boldsymbol{\Sigma} \succeq 0} d(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}^{plug}) \tag{4.6}$$

where $d$ measures the distance between two matrices. For instance, the Frobenius norm (Katayama et al. 2018; Wang et al. 2014) and the element-wise maximum norm (Loh and Tan 2018) are used previously. Then, the nearest matrix $\hat{\boldsymbol{\Sigma}}^{psd}$ would be put into the subsequent multivariate analyses (e.g. the graphical lasso) without modification in the current implementations. However, solving the problem (4.6) comes at the price of such

convenience. When the Frobenius norm is used, (4.6) amounts to a well-known projection onto the convex cone of PSD matrices. The solution can be explicitly expressed by

$$\hat{\boldsymbol{\Sigma}}^{psd} = \boldsymbol{V}\boldsymbol{W}_{+}\boldsymbol{V}^{\mathrm{T}}, \quad \boldsymbol{W}_{+} = \max(\boldsymbol{W}, \boldsymbol{0})$$

where $\hat{\boldsymbol{\Sigma}}^{plug}$ has the spectral decomposition $\boldsymbol{V}\boldsymbol{W}\boldsymbol{V}^{\mathrm{T}}$ and the maximum between two matrices operates element-wisely. The computational cost for this case is mostly from the eigenvalue decomposition. However, the theoretical properties derived for the IPW estimator (e.g. Theorem 1) are not guaranteed to hold for the nearest PSD matrix. In contrast, the convergence rate is preserved when $d$ is the element-wise maximum norm (Loh and Tan 2018) since

$$||\hat{\boldsymbol{\Sigma}}^{psd} - \boldsymbol{\Sigma}||_{max} \leq d(\hat{\boldsymbol{\Sigma}}^{psd}, \hat{\boldsymbol{\Sigma}}^{plug}) + d(\hat{\boldsymbol{\Sigma}}^{plug}, \boldsymbol{\Sigma}) \leq 2d(\hat{\boldsymbol{\Sigma}}^{plug}, \boldsymbol{\Sigma})$$

where the first inequality uses the triangular inequality and the second is from the definition of $\hat{\boldsymbol{\Sigma}}^{psd}$. The algorithm to solve (4.6) with the element-wise maximum norm is first proposed by Xu and Shao (2012) and used in robust covariance estimation context (Han et al. 2014; Loh and Tan 2018). We note, however, by experience that the approximation based on $||\cdot||_{max}$ is computationally heavy so that it often dominates the computation time of multivariate procedures (e.g. the graphical lasso and the CLIME).

# Chapter 5

# Numerical study

In Chapter 5, we perform a number of simulations for estimating a covariance/precision matrix with partially observed data. First, in Chapter 5.2, we experimentally check the convergence rate of the IPW estimator given in our theorems. Next, we investigate the finite sample performance by changing various parameters (e.g. $r = p/n$, missing proportion) in Chapter 5.3. To evaluate estimation accuracy and support recovery of the Gaussian graphical model, different matrix norms and an area under the receiver operating characteristic (ROC) curve are used. In Chapter 5.4, we conduct a comparison study between several imputation methods and the IPW method. Lastly, it is numerically verified in Chapter 5.5 that the coordinate descent algorithm fails when the IPW estimator is plugged-in as discussed in Chapter 4, but Hsieh et al. (2014)'s algorithm does not.

## 5.1 Setting

### 5.1.1 Data generation

We generate Gaussian random vectors $X_i$, $i = 1, \ldots, n$, in $\mathbb{R}^p$ with mean vector 0 and precision matrix $\mathbf{\Omega} = (\omega_{ij}, 1 \leq i, j \leq p)$ under different pairs of $n = 50, 100, 200$ and $p$ satisfying $r(= p/n) = 0.2, 1, 2$. We consider threes types of a precision matrix as follows, which have been used in the previous literature (Cai et al. 2011; Loh and Wainwright 2012).

1. Chain-structured graph : The edge set $E$ of a graph is defined by the structure of a chain graph. $\omega_{ij} = 0.1$, if $(i, j) \in E$, and 0, otherwise; $\omega_{ii} = 1$.

2. Star-structured graph : The edge set $E$ of a graph is defined by the structure of a star-shaped graph. $\omega_{ij} = 0.9/\sqrt{p-1}$[1], if $(i, j) \in E$, and 0, otherwise; $\omega_{ii} = 1$.

3. Erdös-Rényi random graph : Each off-diagonal component in the upper part of $\boldsymbol{B}$ is independently generated, and equals to 0.5 with probability $\log p/p$ and 0 otherwise. Then, the lower part of $\boldsymbol{B}$ is filled with the transposed upper part. Finally, some positive constant is added to the diagonals, i.e., $\mathbf{\Omega} = \boldsymbol{B} + 1.5|\lambda_{min}|\ \mathbf{I}$, to satisfy PDness where $\lambda_{min}$ is the smallest eigenvalue of $\boldsymbol{B}$.

Every $\mathbf{\Omega}$ is rescaled so that the largest eigenvalue of $\mathbf{\Omega}$ is set as 1.

Two structures are under consideration to impose missing on data. The first structure is the independent structure where every component of $X_i$ is

[1]The off-diagonal element $\omega_{ij}$ should be less than $1/\sqrt{p-1}$ to satisfy $\mathbf{\Omega} \succ 0$.

independently exposed to missing with equal probability;

$$\delta_{ik} \sim \text{Ber}(\pi^{(1)}), \quad k = 1, \ldots, p, \text{ independently} \tag{5.1}$$

where $0 < \pi^{(1)} < 1$. Another structure is designed to model dependency within missing indicators. We assume missingness in the first half of $p$ components (assume even $p$ here) forces missing values in the other halves. First, we generate $p$ independent missing indicators as before

$$\tilde{\delta}_{ik} \sim \text{Ber}(\pi^{(2)}), \quad k = 1, \ldots, p, \text{ independently},$$

for $0 < \pi^{(2)} < 1$. Then, dependent indicators are defined by

$$\delta_{ik} = \tilde{\delta}_{ik}, \quad \delta_{i,k+p/2} = \min\{\tilde{\delta}_{ik}, \tilde{\delta}_{i,k+p/2}\}, \quad k = 1, \ldots, p/2.$$

Thus, the $(k + p/2)$-th component cannot be observed unless its pair is observed, or $\delta_{ik} = 1$ $(k = 1, \ldots, p/2)$. An average proportion of missing elements is $1 - \pi^{(1)}$ for the independent case and $(1 - \pi^{(2)})(2 + \pi^{(2)})/2$ for the dependent case. Consequently, the proportion of missing denoted by $\alpha$ can be tuned by changing $\pi^{(1)}$ or $\pi^{(2)}$. For example, under the dependent missing structure, for $\alpha = 0.3$, $\pi^{(2)}$ is uniquely determined by solving the quadratic equation

$$(1 - \pi^{(2)})(2 + \pi^{(2)})/2 = 0.3.$$

We choose different values $\alpha = 0, 0.15, 0.3$. The case $\alpha = 0$ where all samples are completely observed is included as a reference.

We generate 10 data sets for each scenario to capture variability from randomness.

### 5.1.2 Estimators

We compare two types of a plug-in estimator: $\hat{\mathbf{\Sigma}}^{IPW}$, an oracle type estimator labeled by "orc" and $\hat{\mathbf{\Sigma}}^{emp}$, an empirical type estimator labeled by "emp". A closed form of the weight $\pi_{k\ell}$ is accessible according to each missing structure, so the oracle IPW estimator is explicitly computable. It is noteworthy that the estimator $\hat{\mathbf{\Sigma}}^{emp}$ is used in Kolar and Xing (2012), but their theoretical analysis is limited to the independent missing structure.

Based on our experience, the graphical lasso is preferred to the CLIME in estimation of sparse precision matrices since the implemented R packages are either too conservative to find true edges (R package `fastclime`) or too slow (R package `clime`). We exploit `QUIC` algorithm proposed by Hsieh et al. (2014) to solve the graphical lasso (4.1). The grid of a tuning parameter $\lambda \in \Lambda$ is defined adaptively to the plug-in matrix $\hat{\mathbf{\Sigma}}^{plug}$

$$\Lambda = \Big\{ \exp\{\log(\kappa M) - d\log(\kappa)/(T-1)\} : d = 0, \ldots, T-1 \Big\},$$

where $0 < \kappa < 1$ and $M = \big|\big|\hat{\mathbf{\Sigma}}^{plug} - \mathrm{diag}(\hat{\mathbf{\Sigma}}^{plug})\big|\big|_{max}$. Note that the points in $\Lambda$ are equally spaced in log-scale from $\log(\kappa M)$ to $\log M$ by length of $T$. $\kappa$ is set as 0.1 and $T$ as 10.

## 5.2 The rate of convergence

We verify our theoretical results (Theorem 1 and 3) by computing the element-wise maximum deviation $||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_{max}$. We fix $p = 100$ and vary the sample size in $20 \leq n \leq 10000$. We repeat each scenario 20 times and plot the log-transformed empirical distance against $\log(n/p)$. Different plug-in estimators ("orc", "emp") and precision matrices (chain, star, random) are under consideration.

Figure 5.1 shows that each graph connecting the averaged distances nearly forms a straight line. The results in the column "orc" confirm the rate of convergence in Theorem 1, while those in another column "emp" confirms that in Theorem 3.

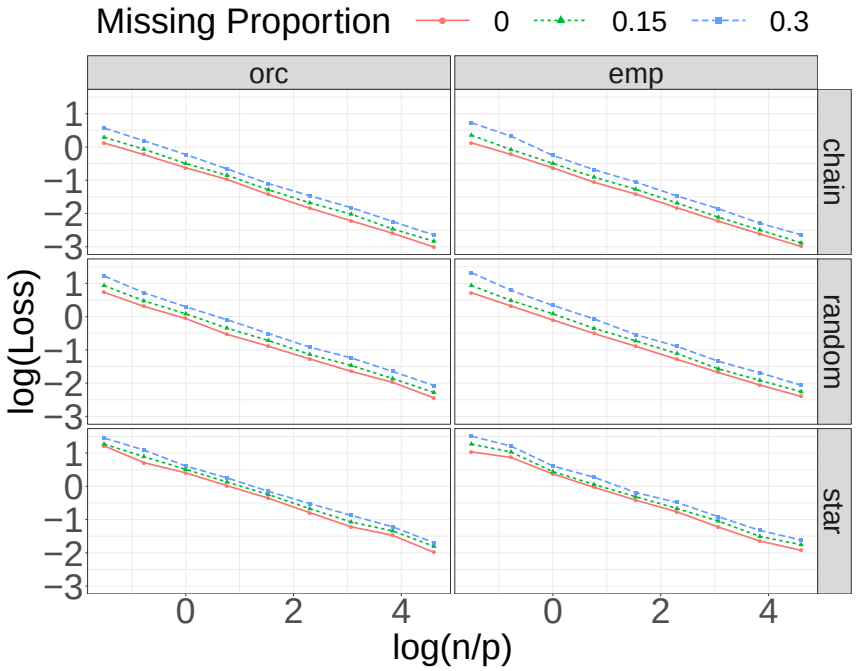


Figure 5.1: Convergence rate of the plug-in matrix ("orc"=$\hat{\mathbf{\Sigma}}^{IPW}$, "emp"=$\hat{\mathbf{\Sigma}}^{emp}$) against $\log(n/p)$. Loss is computed by the element-wise maximum norm between the plug-in matrix and the true covariance matrix. The dependent missing structure and $p = 100$ are assumed. Each dot (or mark) is an average loss from 20 repetitions.

## 5.3 Performance comparison

### 5.3.1 Estimation accuracy

We numerically examine behaviors of the inverse covariance matrix estimated using the IPW estimator as varying simulation parameters. To this end, the Frobenius and spectral norms are used to measure a distance of an estimator. We fix the $\lfloor 0.7\,T \rfloor$-th tuning parameter in $\Lambda$ (in increasing order) to get a single sparse precision matrix, because selection of the tuning parameter is not of our primary interest and our findings stated below does not change much according to the tuning parameter.

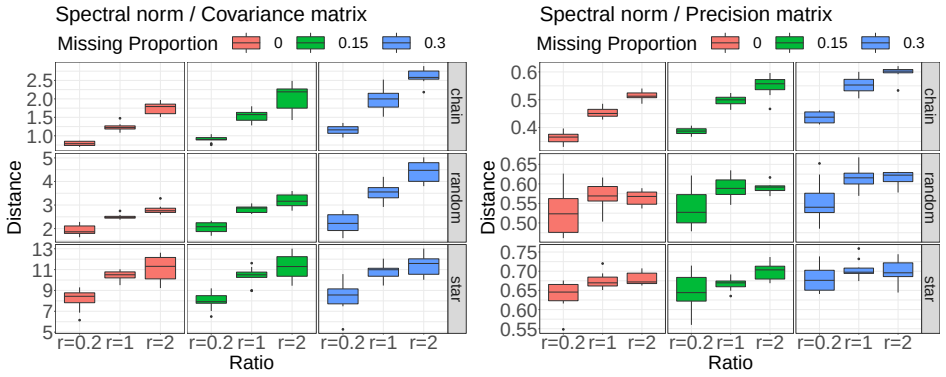


Figure 5.2: Boxplots of the spectral norm with different ratios $r(=p/n) = 0.2, 1, 2$. The dependent missing structure and $n = 100$ are assumed. The oracle IPW estimator is plugged-in. $||\hat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}||$ (left) and $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||$ (right) are measured.

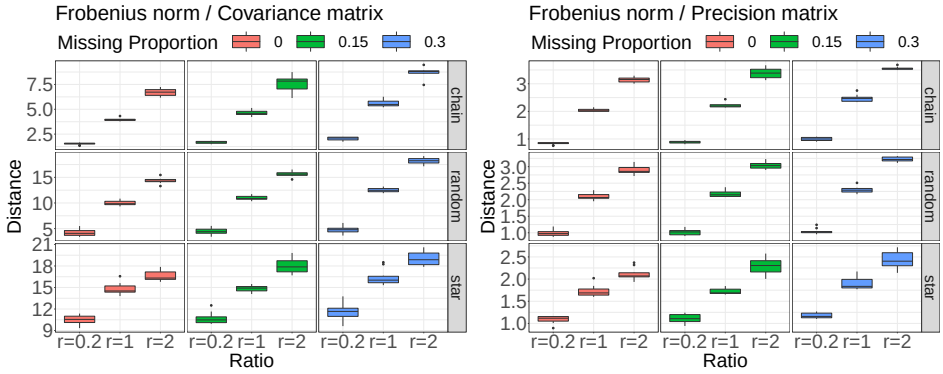


Figure 5.3: Boxplots of the Frobenius norm with different ratios $r(= p/n) = 0.2, 1, 2$. The dependent missing structure and $n = 100$ are assumed. The oracle IPW estimator is plugged-in. $||\hat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}||$ (left) and $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||$ (right) are measured.

Figure 5.2 and 5.3 show that the ratio of the sample size and dimension is one of the key factors that determines the magnitude of estimation error. It is uniformly observed that larger size of a precision matrix is more difficult to estimate, but the degree of difficulty depends on the shape of the true graphs (or precision matrix).

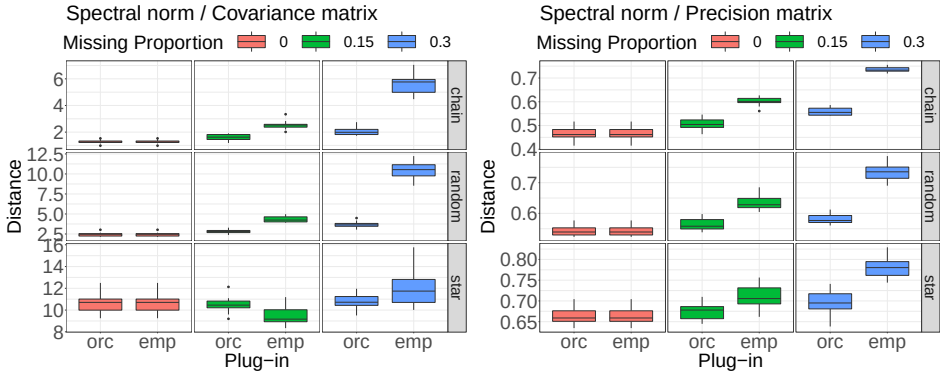


Figure 5.4: Boxplots of the spectral norm with different plug-in estimators ("emp" and "orc"). The dependent missing structure, $n = 100$ and $r = 1$ are assumed. $||\hat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}||$ (left) and $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||$ (right) are measured.

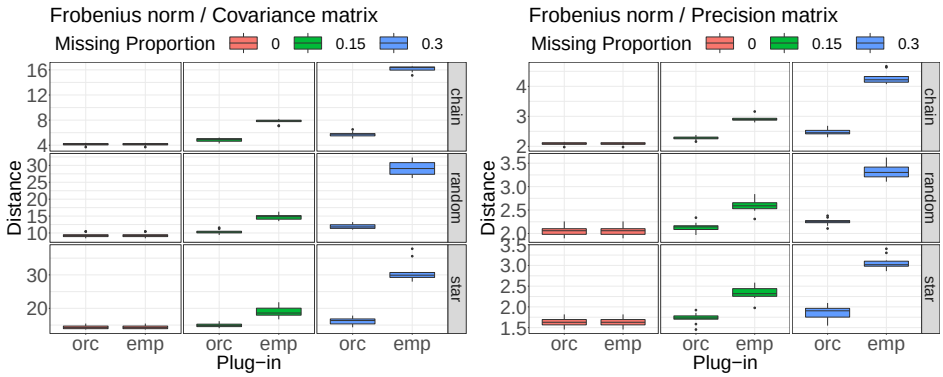


Figure 5.5: Boxplots of the Frobenius norm with different plug-in estimators ("emp" and "orc"). The dependent missing structure, $n = 100$ and $r = 1$ are assumed. $||\hat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}||$ (left) and $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||$ (right) are measured.

Figure 5.4 and 5.5 compare the performance of the two plug-in matrices. When complete data is available, no adjustment for missing is needed so that there is no difference in errors (see the leftmost red boxplots in each sub-figure). If missing occurs in data, the precision matrix estimator based

on the oracle IPW estimator is closer to the true matrix (either $\mathbf{\Sigma}$ or $\mathbf{\Omega}$), and the extent is more evident as the missing proportion $\alpha$ increases.

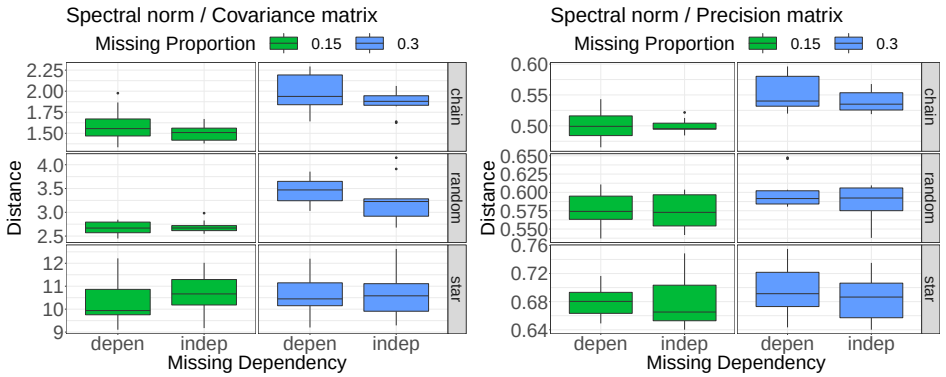


Figure 5.6: Boxplots of the spectral norm with different missing structures ("depen" and "indep"). $n = 100$ and $r = 1$ are assumed. The oracle IPW estimator is plugged-in. $||\hat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}||$ (left) and $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||$ (right) are measured.

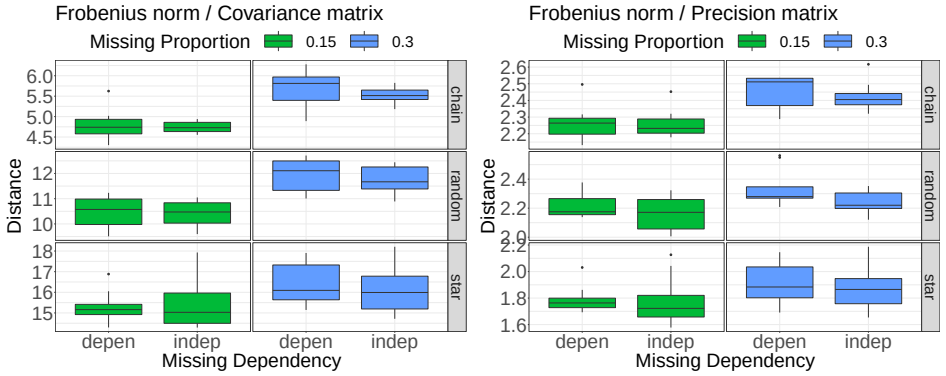


Figure 5.7: Boxplots of the Frobenius norm with different missing structures ("depen" and "indep"). $n = 100$ and $r = 1$ are assumed. The oracle IPW estimator is plugged-in. $||\hat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}||$ (left) and $||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||$ (right) are measured.

Figure 5.6 and 5.7 imply that dependency in missing degrades estimation accuracy, as the missing proportion is set at the same level in both missing structures. We do not show the results when using complete data (i.e., $\alpha = 0$) since the two missing structures are the same by definition.

### 5.3.2 Support recovery

We investigate the support recovery of the Gaussian graphical model using the ROC curve. It is observed that the ROC curves end at different false positive rate (FPR) values, especially when different missing proportions are assumed (see Figure 5.8).

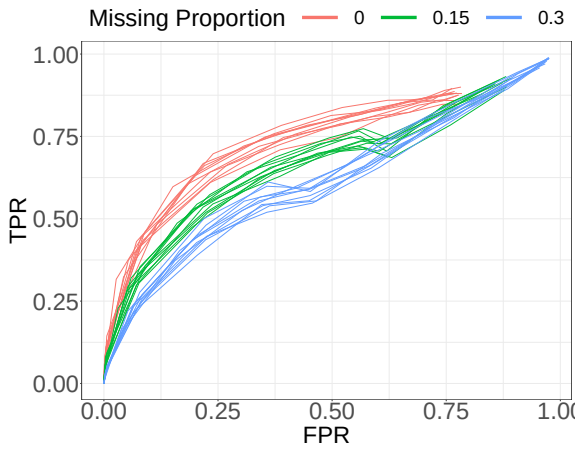


Figure 5.8: The ROC curves according to different missing proportions with 10 times of repetition. $n = 100$, $r = 1$, a random graph structure, and the dependent missing structure are assumed. The oracle IPW estimator is plugged-in.

Thus, it is not fair to directly compare an area under the curve (AUC) because the maximum value of AUC depends on the endpoint (largest value)

of FPR and thus cannot reach 1 if the endpoint is less than 1. Instead, we use the rescaled partial AUC (pAUC) proposed by Walter (2005). The pAUC rescales the AUC by the largest FPR in the ROC curve (see Walter (2005) for more details). Then, the rescaled AUCs from different curves that end at different FPR values have the same range $[0, 1]$.

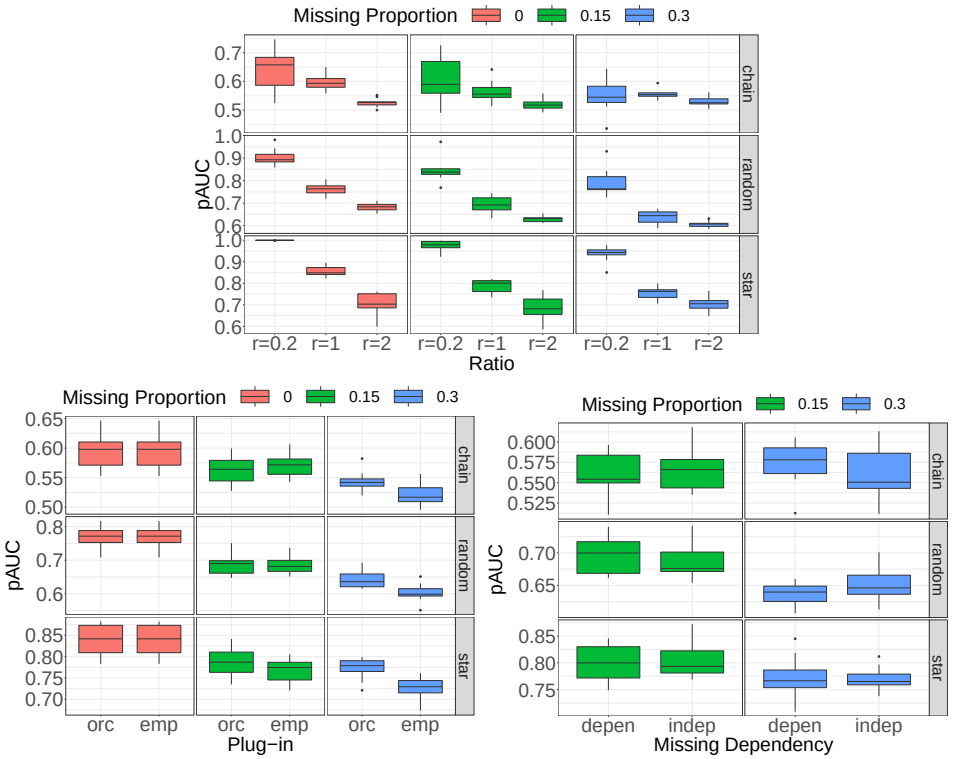


Figure 5.9: (Top) Boxplots of the pAUC with different ratios $r(= p/n) = 0.2, 1, 2$. Dependent missing structure and $n = 100$ are assumed. The oracle IPW estimator is plugged-in. (Bottom left) Boxplots of the pAUC for support recovery with different plug-in estimators. $n = 100$, $r = 2$, and the dependent missing structure are assumed. (Bottom right) Boxplots of the pAUC for support recovery with different missing structures ("depen" and "indep"). $n = 100$ and $r = 1$ are assumed. The oracle IPW estimator is plugged-in.

Figure 5.9 shows the results of the pAUC as the simulation parameters are varying. Considering a large value of the pAUC implies better performance in the support recovery, we have similar interpretations based on the given

results as before.

## 5.4 Comparison with imputation methods

In the missing data context, unobserved data is often substituted by some function of observed values. One very intuitive way to do it is the imputation method. Once the pseudo complete data is produced, we perform a usual statistical analysis. In this experiment, we compare different (single) imputation approaches with the IPW estimator for the precision matrix estimation.

Imputation methods we use are "median" (a median of available data for each variable), "pmm" (predictive mean matching from R package `Hmisc` (Harrell Jr et al. 2019)), "knn" (an average of k-nearest neighbors from R package `impute` (Hastie et al. 2018)), "cart", "rf", and "norm" (regression-based methods from R package `mice` (van Buuren and Groothuis-Oudshoorn 2011)). We use the default parameter setting for each R function. More details of each method can be found in each reference.

By fixing $n = 100$ and $r = 1, 2$, we generate 10 random data sets based on different precision matrices. Missing observations are produced under the independent structure. Once missing observations are filled by a single imputation method, then we compute the sample covariance matrix with the imputed complete data and carry out the precision matrix estimation using the QUIC algorithm. We compare the competing methods based on support recovery of the estimated precision matrix. Figure 5.10 shows the pAUC values, where the IPW method using the empirical estimator ("emp") achieves the largest pAUC compared to the imputation approaches. This is more distinct when the dimension is larger than the

sample size (i.e., $r = 2$). The results demonstrate that the IPW method is not only theoretically solid, but also practically useful. Admittedly, we have not thoroughly examined more diverse and complex imputation methods that may produce better performance, which calls for extensive numerical studies in upcoming works.

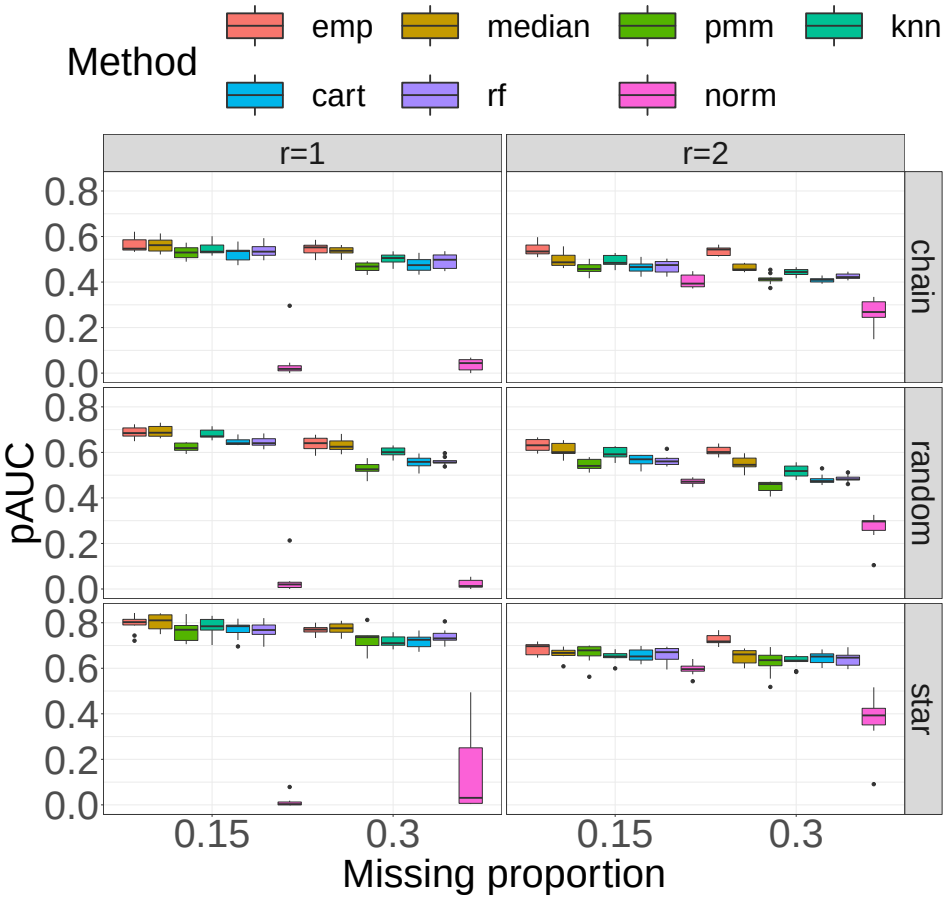


Figure 5.10: Comparison of the pAUC values for different approaches to handle missing in estimating a sparse precision matrix. $r = 1, 2$, $n = 100$, and the independent missing structure are assumed. The empirical IPW estimator is plugged-in. 10 random data sets are used.

## 5.5 Failure of Algorithm 1 under missing data

It is mentioned that the undesirable property, non-PSDness, of the IPW estimator may hamper downstream multivariate procedures. We give one of

the examples where it causes a problem; the graphical lasso. Recall that the existing algorithms available in `glasso` and `huge` packages are not suitable especially with the tuning parameter fixed at small $\lambda$, since they use the non-PSD initial matrix $\mathbf{\Sigma}^{(0)} = \hat{\mathbf{\Sigma}}^{IPW} + \lambda\mathbf{I}$. As a consequence, in Figure 5.11, the blue solid ROC curves end at FPR values far less than 1 when the coordinate descent algorithm provided in `huge` is used. On the contrary, the QUIC algorithm (red dashed) returns a full length of ROC curves. It is noted that since the graphical lasso has a unique solution, two algorithms create the same path, as long as convergent.

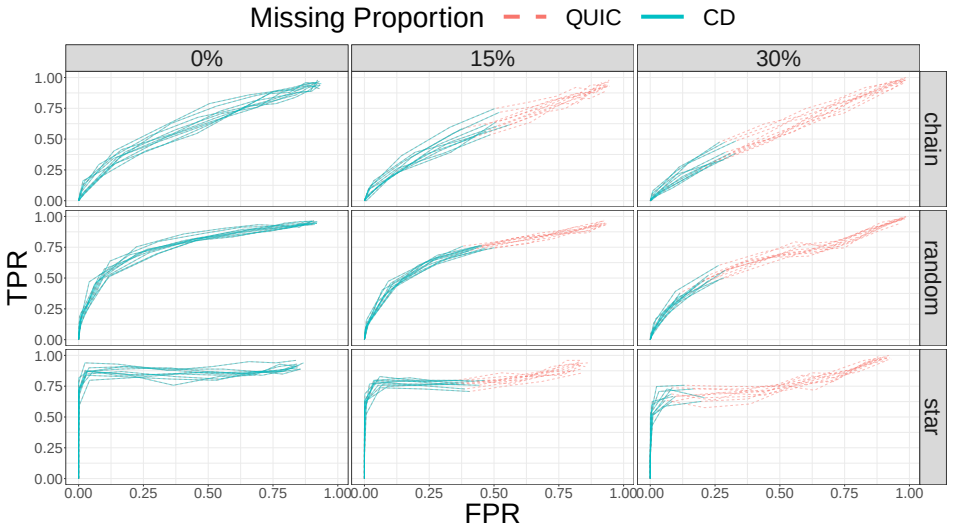


Figure 5.11: Comparison of ROC curves between two different algorithms for solving the graphical lasso using incomplete data. $n = 100$, $r = 1$, and the dependent missing structure are assumed. The oracle IPW estimator is plugged-in. 10 random data sets are used.

# Chapter 6

# Application to real data

We examine the estimation performance of the IPW estimator through a real data application. We use the riboflavin data available from the R package `hdi`, where 4088 gene expressions are observed across 71 samples. Each variable is log-transformed and then centered. We select 1000 genes with the largest empirical variances for the sake of simplicity. As in the previous analyses, the QUIC algorithm is used to solve the graphical lasso.

With the complete data set, we solve the graphical lasso (4.1) with a fixed $\lambda$ and set the obtained estimate $\mathbf{\Omega}_\lambda$ as the ground truth precision matrix. We generate three different models with $\lambda_1 < \lambda_2 < \lambda_3$. Note that the estimated precision matrix with a smaller tuning parameter (e.g. $\lambda_1$) gives a denser true model. We also consider another ground-truth precision matrix with an optimal tuning parameter that is chosen by the cross-validation procedure, following Kolar and Xing (2012). Let an index set of $n$ samples split into $K$ folds $\{G_k\}_{k=1}^K$ of equal size. Without samples in the $k$-th fold, we estimate the precision matrix at a fixed $\lambda$, denoted by $\mathbf{\Omega}_\lambda^{(k)}$. We finally choose $\lambda_{CV}$ among a grid of $\lambda$'s that minimizes the cross-validated

(negative) log-likelihood function below;

$$CV(\lambda) = \sum_{k=1}^{K} \sum_{i \in G_k} \Big\{ \log |\mathbf{\Omega}_{\lambda}^{(k)}| + X_i^{\mathrm{T}} \mathbf{\Omega}_{\lambda}^{(k)} X_i \Big\}.$$

We let $\mathbf{\Omega}_{CV} = \mathbf{\Omega}_{\lambda_{CV}}$ the precision matrix at this level of the optimal sparsity $\lambda_{CV}$. It turns out $\lambda_{CV}$ is close to, but slightly smaller than the smallest tuning parameter $\lambda_1$. The four precision matrix models have $36,170$ ($\lambda_1$), $5,860$ ($\lambda_2$), $14$ ($\lambda_3$), $35,630$ ($\lambda_{CV}$) non-zero elements (except diagonals) in each.

We impose missing values on the complete data matrix in a similar manner described in Chapter 5. For this analysis, we assume the independent missing structure and note that results do not alter significantly using the dependent structure. To estimate $\mathbf{\Omega}_{\lambda_i}$, we solve the graphical lasso (4.1) using the incomplete data with the tuning parameter fixed at $\lambda_i$. Since the optimality of the tuning parameter can vary as different data is available due to missing, the cross-validation procedure is separately performed, instead of using the same $\lambda_{CV}$ to estimate $\mathbf{\Omega}_{CV}$. Let $\hat{\mathbf{\Omega}}_{\lambda}^{(k)}$ be the solution with the tuning parameter $\lambda$ without the $k$-th fold of incomplete data. Then, the (cross-validated) log-likelihood is computed over observed data as follows;

$$CV_{mis}(\lambda) = \sum_{k=1}^{K} \sum_{i \in G_k} \Big\{ \log |(\boldsymbol{Q}_i^{(k)})^{-1}| + X_{i,obs}^{\mathrm{T}} (\boldsymbol{Q}_i^{(k)})^{-1} X_{i,obs} \Big\}$$

where $\boldsymbol{Q}_i^{(k)} = ((\hat{\mathbf{\Omega}}_{\lambda}^{(k)})^{-1})_{i,obs} = \Big(((\hat{\mathbf{\Omega}}_{\lambda}^{(k)})^{-1})_{k\ell}, k, \ell \in \{k : \delta_{ik} = 1\}\Big)$ and $X_{i,obs} = \big(X_{ik}, k \in \{k : \delta_{ik} = 1\}\big)^{\mathrm{T}}$. Let $\hat{\lambda}_{CV}$ the optimal parameter that minimizes $CV_{mis}$ and $\hat{\mathbf{\Omega}}_{CV}$ the graphical lasso solution using all observed data at $\hat{\lambda}_{CV}$.

The following Figure 6.1 presents three different measures to access precision matrix estimation. An error distance between the truth and an

estimate is evaluated by the spectral norm. Due to readability, the boxplots of the distance for dense models ("D" and "CV") under the missing proportion 30% are not shown, but their summary statistics are provided in Table 6.1. It is confirmed again that having more missing values yields worse estimation. Also, it is possible to see that the denser model is more difficult to achieve satisfactory accuracy in estimation and graph recovery.

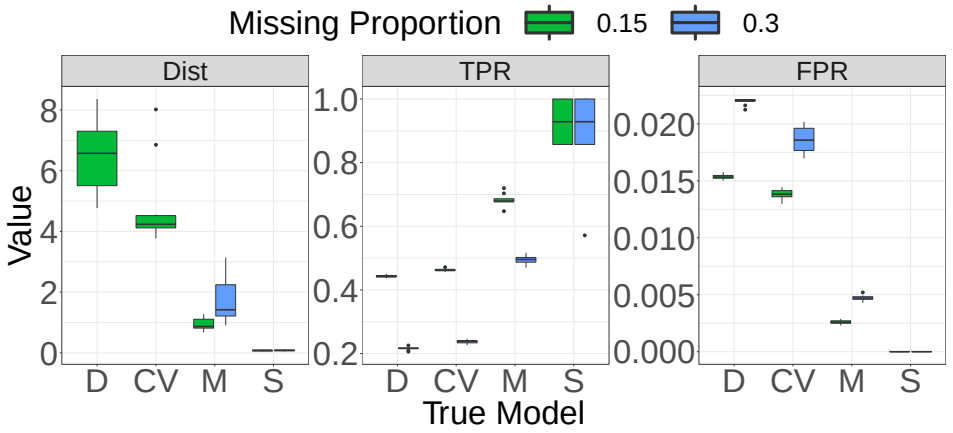


Figure 6.1: Boxplot of performance measures (left: the error distance, middle: TPR, right: FPR) using the riboflavin data. "D", "M", "S", and "CV" on the x-axis stand for the dense ($\lambda_1$), moderate ($\lambda_2$), sparse ($\lambda_3$), and cross-validated ($\lambda_{CV}$) models, respectively. Due to readability, two boxplots for the distance from "D" and "CV" are not shown when the missing proportion is 30%.

| | min | Q1 | Q2 | Q3 | max |
|---|---|---|---|---|---|
| D | 62.135 | 771.178 | 4340.741 | 8749.103 | 16449.95 |
| CV | 26.656 | 30.359 | 53.212 | 3939.772 | 34043.44 |

Table 6.1: Quantiles for the spectral norms of the dense ("D") and cross-validated ("CV") models with the missing proportion 30%.

# Chapter 7

# Discussion

This thesis considers a theoretical establishment of the IPW estimator with missing observations. Contrary to the previous literature, we generalize dependency among missingness, meaning that missing indicators are not necessarily independent across variables. The rate of convergence of the IPW estimator is derived based on the element-wise maximum norm, which is (asymptotically) in the same order of the rate claimed in the past works. Our analysis can be applied to an estimation of a sparse precision matrix. Due to the meta-theorem, the favorable properties (consistency, support recovery) of the final estimator are preserved in missing data context.

The plug-in estimators (e.g. the sample covariance matrix and the IPW estimator) and their concentration are often not of primary interest, but the ultimate goal lies in applying them to downstream procedures (e.g. Hotelling's $\mathrm{T}^2$, a portfolio optimization, etc). In the portfolio optimization, Fan et al. (2012) show that the risk inequality is bounded by the error of the plug-in estimator $\hat{\mathbf{\Sigma}}^{plug}$;

$$|w^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{plug}w - w^{\mathrm{T}}\mathbf{\Sigma}w| \leq ||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_{max}.$$

Here, $w$ and $\mathbf{\Sigma}$ are true (or optimal) parameters. However, it is still elusive how the rate $||\hat{w} - w||_V$ for the optimal solution $\hat{w}$ that minimizes the risk $t \mapsto t^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{plug}t$ is linked to the rate $||\hat{\mathbf{\Sigma}}^{plug} - \mathbf{\Sigma}||_M$ of the plug-in estimator. $||\cdot||_V$ and $||\cdot||_M$ are some norms of a vector and a matrix, respectively. This line of research could be interesting for future work and in urgent need, not to mention its extension to the missing data context.

The underlying assumptions on the missing mechanism (i.e., MCAR) and the missing structure (i.e., identical dependency across samples) are essentially not verifiable, but it is natural to think of extending our results to the cases beyond such patterns. Recall from the text below the equation (2.4) that the IPW estimator under the general missing mechanism is to be defined by

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{R}_i * X_i X_i^{\mathrm{T}}$$

where $\boldsymbol{R}_i = \big(\delta_{ij}\delta_{ik}/\pi_{i,jk}, 1 \le j,k \le p\big)$ and $\pi_{i,jk} = \mathrm{P}(\delta_{ij}\delta_{ik} = 1 | X_i, W_i)$ with external factors $W_i$. It is easy to show the above estimator is still unbiased for $\mathbf{\Sigma}$. Now, let us consider the missing at random (MAR) assumption (additionally assume $W_i = \emptyset$), i.e.,

**Assumption 4** (Missing at random)**.** *An event that an observation is missing is independent with unobserved random variables given observed variables.*

which essentially says independence between $\delta_i$ and $X_i$ given observed data $X_{i,obs}$. In this case, it is not straightforward to follow the analysis in this thesis. For example, one should work on the calculation displayed below;

$$\mathbb{E}\left[\frac{\delta_{ik}X_{ik}^2}{\pi_{i,k\ell}}\right] = \mathbb{E}\left[X_{ik}^2\mathbb{E}\left[\frac{\delta_{ik}}{\pi_{i,k\ell}}\middle| X_{i,obs}\right]\right] = \mathbb{E}\left[\frac{\pi_{i,k}X_{ik}^2}{\pi_{i,k\ell}}\right]. \tag{7.1}$$

Unfortunately, the fraction $\pi_{i,k}/\pi_{i,k\ell}$ cannot be out of expectation since $\pi_{i,k}, \pi_{i,k\ell}$ are functions of $X_{i,obs}$, which makes it difficult to explicitly express (7.1) in terms of $\sigma_{kk}$ and others. This was possible under MCAR because of the independence of $\pi_{i,k}$ from $X_i$. It would be interesting to identify suitable assumptions that are less stronger than MCAR, but still guarantee the missing probability to be free from $X_{i,obs}$.

# Bibliography

Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Boucheron, S., Lugosi, G., and Pascal, M. (2016). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

Cai, T., Liu, W., and Luo, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-

dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59.

Cai, T. T. and Zhang, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis*, 150:55–74.

Cui, R., Groot, P., and Heskes, T. (2017). Robust estimation of gaussian copula causal structure from mixed data with missing values. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 835–840.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Glanz, H. and Carvalho, L. (2018). An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. *Journal of Multivariate Analysis*, 167:31–48.

Han, F., Lu, J., and Liu, H. (2014). Robust scatter matrix estimation for high dimensional distributions with heavy tails. *Technical report, Princeton University.*

Harrell Jr, F. E., with contributions from Charles Dupont, and many others. (2019). *Hmisc: Harrell Miscellaneous.* R package version 4.2-0.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2018). *impute: impute: Imputation for microarray data*. R package version 1.56.0.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014). Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15:2911–2947.

Huang, J. Z., Liu, L., and Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics*, 16(1):189–209.

Katayama, S., Fujisawa, H., and Drton, M. (2018). Robust and sparse gaussian graphical modelling under cell-wise contamination. *Stat*, 7(1):e181.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, page 10. Chapman and Hall, 1 edition.

Kolar, M. and Xing, E. P. (2012). Estimating sparse precision matrices from data with missing values. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 635–642, USA. Omnipress.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An imputation–regularized optimization algorithm for high dimensional missing data problems and beyond. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):899–926.

Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Loh, P.-L. and Tan, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under $\epsilon$-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Pang, H., Liu, H., and Vanderbei, R. (2014). The fastclime package for linear programming and large-scale precision matrix estimation in r. *Journal of Machine Learning Research*, 15(1):489–493.

Park, S. and Lim, J. (2019). Non-asymptotic rate for high-dimensional covariance estimation with non-independent missing observations. *Statistics & Probability Letters*, 153:113–123.

Pavez, E. and Ortega, A. (2019). Covariance matrix estimation with non uniform and data dependent missing observations.

Rao, M., Javidi, T., Eldar, Y. C., and Goldsmith, A. (2017). Estimation in autoregressive processes with partial observations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4212–4216.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Rudelson, M. and Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:9 pp.

Saulis, L. and Statulevičius, V. (1991). *Limit theorems for large deviations*. Springer Science Business Media.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871.

Städler, N. and Bühlmann, P. (2012). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235.

Thai, J., Hunter, T., Akametalu, A. K., Tomlin, C. J., and Bayen, A. M. (2014). Inverse covariance estimation from data with missing values using the concave-convex procedure. In *53rd IEEE Conference on Decision and Control*, pages 5736–5742.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, page 11–37. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Walter, S. D. (2005). The partial area under the summary roc curve. *Statistics in Medicine*, 24(13):2025–2040.

Wang, H., Fazayeli, F., Chatterjee, S., and Banerjee, A. (2014). Gaussian Copula Precision Estimation with Missing Values. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 978–986, Reykjavik, Iceland. PMLR.

Xu, M. H. and Shao, H. (2012). Solving the matrix nearness problem in the maximum norm by applying a projection and contraction method. *Advances in Operations Research*, 2012:1–15.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13:1059–1062.

# 국문초록

결측이 없는 자료에서 표본 공분산 행렬 $\boldsymbol{S}$은 다양한 다변량 통계 절차를 개시하는 핵심 통계량이다. 가령, 구조를 가진 (역)공분산 행렬 추정, 주성분 분석, 그래프 모형 등에 $\boldsymbol{S}$가 사용된다. 반면, 결측 자료를 이용하여 계산한 표본 공분산 행렬은 편향되어 있어 바람직하지 못하다. 기존 연구에서는 이러한 편향을 수정해주기 위해 역확률 가중치(IPW라 표기함)라는 간단한 보정 절차를 사용하였으며, 이를 통해 IPW 추정량을 제안하였다. IPW 통계량은 결측이 있는 자료에서 기존의 표본 공분산 행렬의 역할을 대신하며 기성 다변량 절차에 삽입하는 식으로 이용되어 왔다. 하지만, 이 추정량의 이론적 성질 - 예를 들어 집중 부등식 - 은 아주 단순한 구조의 결측 구조(모든 변수가 독립적으고 같은 확률로 결측에 노출이 됨) 하에서만 연구되어 왔다.

이에 본 학위 논문에서는 일반적인 결측 구조 하에서 발생한 결측 자료를 이용하여 계산한 IPW 추정량의 편차를 연구하고자 한다. 본 논문에서는 IPW 추정량의 원소별 최댓값 행렬 노음에 기반한 최적 수렴 속도 $O_p(\sqrt{\log p/n})$를 증명한다. 또한 암묵적인 가정들(평균 그리고/혹은 결측 확률을 알고 있음)을 완화하여 유사한 편차 부등식을 유도한다. 유도된 최적의 수렴 속도는 특히 역공분산 행렬 추정에 중대한 의미를 갖고 있다. 이는 IPW 추정량의 속도가 최종 역공분산 행렬 추정량의 속도를 지배한다는 "메타 정리"(Liu et al. 2012)에 의해 뒷받침 된다. 모의 실험 연구에서는 IPW 추정량이 양의 준정부호 성질을 만족하지 않는 것에 대해 논하고, 대치법을 이용한 추정량과의 비교를 다루고 있다. 이는 실용적인 측면에서 중요한 논의들이다.

**주요어:** 수렴 속도, 공분산 행렬, 의존적 결측 구조, 역확률 가중치, 결측 자료.

**학　번:** 2013 – 22899